WE EXCHANGE DATA ALL THE TIME

A case study on data conceptualisation and data sharing by researchers in the context of open data policies

Madeleine Dutoit

OSLOMET

Preface

I would like to thank the many fellow beings that in their own ways have contributed to make this thesis possible, but first just a couple of reflections on the work process itself.

When I studied at art schools many years ago, the teachers repeatedly told us that learning to draw was all about learning how *to see*. After much practice, observing and reobserving, we slowly learned to perceive the numerous colour nuances of surfaces as well as the form of spaces that surrounded objects. I have thought of this many times during the work with this thesis. To learn how to carry out a research project, including countless steps and decisions, has in many ways been to open eyes and mind to things I could not previously perceive. It is about developing a sensitivity for nuances, relationships, and conditions beyond the well-known and apparent. Like art, to develop an investigation is also a very creative process. Blank pages are slowly filled with traces of a discovery process that did not exist previously, and in this exploration, nothing is right or wrong as long as you can defend your choices, and as long as there is "cognitive harmony", as one of my supervisors said. Like many have told me along the way, after pursuing a PhD programme one will never be the same again.

Before I started as a doctoral student, it was impossible to imagine the challenges that awaited, how surprising, fascinating, and constantly transforming daily work would be; at times working was like walking in dense fog, being unclear and uncertain of where to place the foot next. The thesis text is just a very small trace of several years of work and intense learning. Some parts of the monography took months or years before materialising, it was an iterative process, pending back and forth like the movements of a 3D printer, each time leaving a thin layer of deepened insight. In the end, a creation, a monography, took form.

I would first like to thank my supervisors heartily, Professor Katriina Byström at Oslo Metropolitan University and Professor Jutta Haider at the University of Borås, to whom I am infinitely grateful. A research education is in my opinion a true example of *learning by doing* and you both stood firmly by my side as companions during this adventure, guiding me onwards with continuing support and encouragement. Due to our many discussions, when you demonstrated your profound knowledge and experience, you helped me gradually develop my own scientific reasoning. With empathy, generosity, and humour you also showed the many aspects of what is means to live and work as a researcher. I hope our paths will cross again. To be assumed at Oslomet has meant becoming invited to develop my acquaintance with Norwegian people and culture. Getting to know Norway better has been a joyful experience,

Abstract

This dissertation places itself within the area of Scholarly communication research of Library and Information Science. It investigates the relationship between researcher and research data and researcher and data sharing from a researcher perspective. The purpose of this thesis is to enrich our understanding of researchers' data sharing in the context of the open data initiatives of external authorities.

To make research data findable and accessible is a part of the ongoing open science movement. Accessible and reusable research data are stated to increase possibilities to resolve societal challenges and strengthen competitiveness. Stakeholders within the scholarly communication ecosystems use data policies to steer how researchers share research data. Researchers must therefore increasingly respond to requirements about making generated research data accessible. The new data sharing initiatives have potentially comprehensive implications for how research will be carried out in the future.

In this case study, an interdisciplinary, international research group of 18 researchers within the STEM disciplines was investigated. The researchers encountered a data policy via their Horizon 2020 funding and had to make research data accessible in a research data repository and develop a data management plan. The empirical material consists of transcribed interviews, observation notes, and documents that were collected over the period of over a year. Wenger's community of practice was used as conceptual framework to direct attention to the researchers' shared perspectives of data practices. The study's aim was to elucidate the researchers' conceptualisations of research data and the ways in which data sharing is an element interwoven in various practices arising from participation in an interdisciplinary community that is bound by a data policy. Three specific research questions were posed. 1: How are data negotiated and reproduced within the group? What are data to the researchers, and when? 2: How do researchers mutually account for the data policy? 3: How do the views on research data and data policies relate to how researchers aim to share data? How can data sharing be imagined as a constitutive yet negotiable element in interdisciplinary research practices?

The results showed an unexpected complexity regarding the researchers' perspectives on data. They shared a definite conceptualisation of data as experimental results and measurements. In parallel the term *data* was elastic and used inclusively for several types of information.

Significant differences in terminology use and material representations of data were found between theoreticians and experimentalists. The understanding of the concept was not discussed and appeared to be anchored in the researchers' training.

Notwithstanding the researchers' well-developed skills in data sharing, the encounter with the policy revealed a lack of knowledge necessary for being able to respond to the data policy. A learning process was initiated while the group tried to avoid changing their existing data sharing practices. The researchers viewed the policy's indicated data sharing methods as meaningless, as opposed to their existing data sharing practices. The essential meanings of data sharing were embedded within what were seen as important activities, or not; how data were shared with others was deeply anchored within well motivated routines developed to meet their needs. Because of these understandings, together with fear that outsiders would not understand the data, competing claims, and lack of a suitable repository, the policy had few implications for how the researchers aimed to share data. The policy's long-term effects should however not be disregarded. The insights of this thesis are valuable because policymakers, research funders, developers and providers of academic support and librarians need to be able to relate to researchers' views on data and data sharing when developing appropriate directives and satisfactory research services during the transition to open science.

Sammanfattning (Swedish summary)

Denna avhandling placerar sig inom området för Vetenskaplig kommunikation i ämnet Biblioteks- och informationsvetenskap. Fallstudien undersöker relationen mellan forskare och forskningsdata samt mellan forskare och datadelning utifrån ett forskarperspektiv. Ändamålet är att berika vår nuvarande förståelse för hur forskare gör forskningsdata tillgängliga i en kontext där externa auktoriteter ställer krav på datadelning.

Att forskningsdata görs tillgängliga och användbara utgör en del av den vidare samtida utvecklingen mot ett öppet vetenskapssystem genom vilket forskningen stegvis är tänkt att närma sig det omgivande samhället. Avsikten är att fler skall kunna ta del av och/eller återanvända forskningsdata vilket i ett större perspektiv sägs vara att skapa bättre förutsättningar för att lösa samhällets utmaningar och stärka konkurrenskraften. Inom ekosystemen för vetenskaplig kommunikation använder intressenter datapolicyer som instrument för att styra hur forskare ska dela forskningsdata med andra. Det medför att forskare allt oftare måste svara mot krav från exempelvis forskningsfinansiärer att göra producerade forskningsdata tillgängliga. Dessa krav kan få omfattande följder för hur forskning utförs i framtiden.

I denna studie undersöktes en interdisciplinär, multinationell forskargrupp bestående av 18 forskare inom naturvetenskapliga discipliner. Forskarna mötte för första gången en datapolicy via finansieringen från EU's ramprogram Horizon 2020. De omfattades av krav på att göra forskningsdata tillgängliga i öppna databaser samt att formulera en datahanteringsplan. För studien samlades empiriskt material in genom observationer, intervjuer och dokument under ett drygt år. För att urskilja gruppens gemensamma perspektiv på datapraktiker användes Wengers Community of practice som teoretiskt ramverk. Studiens syfte var att klarlägga forskares förståelser av forskningsdata samt hur datadelning utgör en beståndsdel av olika forskarpraktiker som skapas genom aktiv medverkan i en interdisciplinär gemenskap som måste följa en datapolicy. För att uppnå syftet ställdes specifika frågor. 1: Hur förhandlas och reproduceras forskningsdata inom gruppen? Vad är data för forskarna, och när? 2: Hur utformar forskarna ett gemensamt svar mot datapolicyn? 3: På vilket sätt kan forskarnas syn på data samt datapolicy relateras till hur de avser att dela forskningsdata? Hur kan data utgöra både en konstitutiv och förhandlingsbar komponent i interdisciplinära forskningspraktiker?

Gällande forskarnas perspektiv på vad data är påvisade resultaten en oväntad komplexitet. En samsyn rådde kring en bestämd definition av data såsom mätningar, experimentella resultat och företeelser genererade av maskiner. Samtidigt visade sig termen *data* vara elastisk och användes flexibelt och inkluderande för olika informationstyper i det dagliga arbetet. Mellan de experimenterande och teoretiska forskarna iakttogs väsentliga skillnader gällande användning av terminologi och datans materiella representation. Vad data var för forskarna var oreflekterat och diskuterades inte; förståelsen för begreppet data verkade vara grundad i forskarnas utbildning.

Studien visade att datapolicyn belyste områden där forskarna saknade nödvändig kunskap för att svara mot dess krav, trots stor kunskap om datadelning både inom och utanför projektet. En lärandeprocess initierades samtidigt som gruppen visade stort motstånd till att förändra nuvarande datapraktiker. Den främsta förklaringen till motståndet var att de datadelningsmetoder som policyn indikerade ansågs meningslösa, till skillnad från forskarnas nuvarande datapraktiker. Datadelningens grundläggande betydelser var inbäddade i vad forskarna ansåg vara meningsfulla handlingar eller inte. Hur data delades med andra låg djupt förankrat i välmotiverade rutiner som utvecklats ur deras behov och prioriteringar. Dessa förståelser, tillsammans med rädsla att datan inte kunde förstås av utomstående, konkurrerande krav och avsaknad av passande dataarkiv, medverkade till att datapolicyns krav fick liten betydelse för hur forskarna avsåg dela data med andra. Datapolicyers effektivitet kan därför ifrågasättas även om dess långsiktiga betydelse inte kan bortses ifrån. Avhandlingens insikter är värdefulla för beslutsfattare, forskningsfinansiärer och forskarstödjande professioner såsom bibliotekarier som behöver förstå forskares syn på data och datadelning när anpassande riktlinjer och tillfredsställande stöd utvecklas för övergången mot öppen vetenskap.

Contents

1. Introduction

> [We] exchange data all the time... with so many colleagues, so many people and… We have to look on the compatibility and we have some experience of this so that we know 'OK this data format can be read out of the public licence so we use PDFs or something like this'. But this is what we do all the time, so without data management and with data management... [It] does not affect our daily work so much.[1]

Before becoming a doctoral student, I worked as a librarian at Malmö University library in Sweden. My main task was to support researchers in matters concerning publishing. I informed researchers of the possibility to publish in the continually-growing number of open access journals and helped to navigate among these to find high-quality titles. The university had adopted a new open access policy in 2010 that stated that the researchers had to deposit a copy of their research publications in the university's open archive. As part of my work, I also administered the university funds that granted researchers financial support to enable them to pay the publication costs for publishing in open access journals.

During this period, Scandinavian governments were discussing the establishment of national open access policies. These policies would principally be about research results but would in extension in addition include research data. Research funders started to require researchers that were granted financing to make their generated data openly accessible. From our perspective as librarians, we saw this requirement as evidence that the idea of open access had grown into a larger movement. What had started with promoting the publication of research results in openly accessible journals had now come to include making the research data underlying the results openly available. In the university library, we tried to understand what this entailed for us, and we shared this experience we shared with many libraries in this part of the world, for it seemed that libraries and librarians were expected to play one of several important roles in what was happening. All of a sudden, we were included in realising this goal that came from both governmental and institutional forces, and we had to prepare for completely new tasks by developing new skills in order to be able to help researchers make their collected data openly accessible.

---

[1] The quote is taken from an interview with a researcher (B3).

What can research data be? How do researchers create, use, and store them? These were all questions that we needed to understand. Being used to providing services to all university faculties, we knew from experience that research work in many ways differs among areas; for example, social scientists use different methods and technologies than biologists. Would these differences also be mirrored in how they worked with data, that is, how they collected, analysed, shared, and stored their research data? We suspected this to be the case but did not know in what way, and, more importantly, what did we librarians need to know about these things to be able to support the researchers in the process? These questions made me curious. This sparked an interest to understand what data mean to researchers, how researchers share data with other researchers, and how they experience the external requirements on how they should manage their data, issues that all form the core of this study.

The quote and vignette above reflect several aspects that relate to this thesis' subject. This dissertation focuses on researchers and research data and on how sharing data with other persons is an element interwoven within different research practices. The purpose of this thesis is to enrich the existing understanding of how researchers' data sharing is shaped within the context of open data initiatives of external authorities. Increased knowledge of data sharing and data policy from a researcher perspective is of interest for both Library and Information Science researchers, practitioners, and other professionals involved in realising the transition towards open data systems. It is paramount that policymakers, research funders, developers and providers of academic support and librarians can relate to researchers' views on data and data sharing. A deeper insight into these relationships can contribute to the development of appropriate directives and satisfactory research services. I ask questions concerning how researchers share their data or make them accessible to others, how data are valued and how researchers respond to data policies. To begin with, framing the concept of "data" is not easily accomplished and the word has a somewhat fluid meaning. Examples of contemporary attempts to define or approach data will be explored later in this thesis. Since a working definition is needed for this thesis, without thereby pretending it is an all-comprehensive one, *research data* will be defined – for now – as collected or generated information that is used as evidence for research purposes.

For the purposes of this investigation, the act of *data sharing* is defined as researchers intentionally making their data available for others to use. One might ask, what is interesting

about modern data sharing? Have not researchers always shared their data with other researchers, shown new results to peers, or presented data in research articles? These are all relevant questions and data sharing is nothing new. Researchers have indeed been communicating data in many ways throughout history, either voluntarily or of necessity, either formally or informally. I will elaborate on these issues in detail later in the thesis. However, questions concerning data and data sharing, in general, are currently attracting great interest. Since a result of the advent of digitalisation is that many data are either born digital or may be transformed into digital objects, data can now be moved in ways that they could not be before. This mobility is what is new about contemporary data sharing.

As digital, data become mobile and may be distributed widely on a large scale. Coupled with the common assumption that data can be useful to others, this development has opened many visions and opportunities. For instance, countries all over the world are developing infrastructures to make public data accessible to citizens. Estonia is an example of one of the most digital countries in Europe, an initiative termed e-Estonia, involving extensive collaborations across the public and private sector in the area of digitalization. Personal data are another form of data that is also used and reused in various ways on various platforms. The European Union has recently launched a strategy so that it might benefit from the increasing volume of personal, non-personal, industrial, and public data generated in Europe. The aim is to build an infrastructure for data, a single European data space, that provides access to these data. The stated ambition is to empower both business and the public sector and to "better the lives of all of its citizens" (European Commission, 2020a, p. 25). A concrete example of how personal data are used is found in child welfare services, where linked datasets that provide personal information of families are used to inform decision-making (Redden et al., 2020). Additionally, most of us encounter in our everyday life the General Data Protection Regulation, applied in 2018, intended to reinforce people's control over their personal data (GDPR, 2016).

Within the systems of scholarly communication, the possibilities of information technology development have raised an unprecedented level of interest in how researchers communicate their findings to governments, research funders, learned societies and universities, as well as amongst themselves (Jubb & Shorley, 2013, p. XIII). This current moment is in many respects a transitional period. Scientific results are being disseminated faster than ever in electronic scholarly journals, and new publishers have entered the stage for communicating research, which is challenging old subscription forms with alternative business models and open access articles. For research data that previously often remained unpublished material, this

development means that they can easily be made accessible to varying degrees, ranging from full accessibility to all interested parties to restricted access that is granted only after formal request, obligatory registration, and an embargo period. These trends have potentially broad implications for how research can be carried out in the future.

Involved stakeholders and policy-making bodies (such as national academies and research councils, journal publishers, educators, and the public at large) and researchers themselves have varying motives for encouraging data sharing (Borgman, 2012). In general, for science, making research data openly accessible is believed to open up the possibility of building new research on previous research results, encourage research collaborations, and improve research quality. As early as 2004, arguments were presented in support of the principle that "publicly funded research data should be openly available to the maximum extent possible" (Arzberger et al., 2004, p. 136). This accessibility was expected to promote transparency, innovation and progress.

In a wider perspective, for society at large, reuse of digitally-preserved data is regarded as an efficient and cost-effective use of public funds (Mauthner & Parry, 2013). This efficiency in turn must be viewed as the background for the development of increased research budgets that demand return on investments. Making research data accessible is also expected to speed up innovation, securing the role as strong competitive parts among the world's research investors, and contributing to economic growth. For large stakeholders financing research, such as governments and supranational organisations, openly accessible data is considered part of the broader strides towards the ideal of *open science*. Open science is an umbrella concept used with increasing frequency particularly among policymakers like the EU. The core of the concept entails a new approach to the scientific process that fosters sharing and collaboration through the entire research cycle; barriers are removed for sharing different forms of output, resources, methods, or tools from the research process (FOSTER consortium, n.d.). In addition to open access to research data and open access to publications that have been seen as the cornerstones, open science also includes open source code, open notebooks, open peer review and citizen science.

The promotion of research collaboration is another parallel pressure that characterises and shapes contemporary science. Science has in many research areas become a collaborative enterprise as opposed to an individualistic endeavour, and this form it has increased significantly since World War II (Meadows, 1998). With this trend follows an increase of interdisciplinary research as a means of preserving or achieving the good life in a society that is complex, global, and rapidly innovating (Frodeman et al., 2017). Interdisciplinary

collaborations are considered necessary for resolving societal challenges. As science becomes more data-intensive and collaborative, data sharing becomes a more important activity among researchers (Tenopir et al., 2011). In addition, more frequent evaluations of scientific outcomes, including measuring scientific impact and citations via academic reward systems, is another important feature of the research sector today (Jacob & Hellström, 2018).

Promoting data sharing via data policies

To realise the ideas of what has been called an international agenda around research data management (Cox et al., 2017), instruments for steering researchers towards sharing their data are needed. Actors at various levels are therefore rapidly formulating data policies and rules or recommendations for how research data are to be managed. For instance, data sharing can be made a condition for the awarding of funding. In 2014, a majority of research institutions in Australia, Canada, Germany, Ireland, Netherlands, New Zealand and the United Kingdom had or were soon to have research data management policies in place (Cox et al., 2017). In the United Kingdom in 2016, a concordat on open research data was developed by a multi-stakeholder group to ensure that the research data of the UK research community would be made openly available for use by other people (Higher Education Funder Education for England, 2016). In the United States, the National Science Foundation has since 2011 required all proposals to include a plan for data management and data sharing (Tenopir et al., 2015).

The research design of this thesis is a case study that focuses on a social setting, a research group, where core processes for this investigation occurred. The main reason for choosing this particular group was that it was funded within Horizon 2020, the framework programme for research and innovation of the European Union, and therefore the project is included in the Open Research Data Pilot[2] which aims to "improve and maximise access to and reuse of research data generated by Horizon 2020 projects" (European Commission, n.d.-b). Research projects that are granted funding have to develop a data management plan (DMP) and provide open access to research data if possible (OpenAIRE, 2017). These demands will be further described in chapter three. The EU is one of the most important research funders in Europe: from 2014–2020, Horizon 2020 had a total budget of 77 billion Euros, which made it

---

[2] The Open Research Data Pilot (ORD) was launched in 2014 by the European Commission. For two years, only core research areas of H2020 were included in the pilot. After the revision of the work programme in 2017, the pilot was extended to cover all the thematic areas of Horizon 2020 (European Commission, n.d.-a).

historically the largest framework programme for research and innovation in the EU (European Commission, 2018). In a European context, the funder was pioneering with this pilot initiative and played an important role as a driver of increasing open access to research data. Further, open science is one of the main strategies for reaching the EU's goals for research, within which open access to research data is an included instrument (European Commission, 2016).[3] The background of the choice for taking a direction towards open access was a decision in 2012, when a formal recommendation about access to and preservation of scientific information was commanded:

> Open access to publications and data from publicly funded research should be promoted and access to publications made the general principle for projects funded by the EU research Framework Programmes. (EUR-Lex: Access to European Union law, 2012, p. 39)

Member states were encouraged to "Define clear policies for the dissemination of and open access to research data resulting from publicly-funded research" and develop e-infrastructures that addressed all stages of the data life cycle (EUR-Lex: Access to European Union law, 2012, p. 41). The Commission expected that open access to scientific research data "... enhances data quality, reduces the need for duplication of research, speeds up scientific progress and helps to combat scientific fraud" (EUR-Lex: Access to European Union law, 2012, p. 3). In the subsequent EU research and innovation programme (Horizon Europe 2021–2027), openness continues to play a key role across the programme and open science practices are to be "mainstreamed" (European Commission, 2020b).

The challenge of changing data sharing practices

Having research data policies in place is an important step towards reaching the goals of making data accessible, but policies in themselves do not transform ideals into practices. To implement the ideas of making data openly accessible as resources for various uses has been described as a conundrum. It has been suggested that, underneath the seemingly simple visions of open data, "thick layers of complexity about the nature of data, research, innovation, and scholarship, incentives and rewards, economics and intellectual property, and

---

[3] The Commission defines open access as "the practice of providing on-line access to scientific information that is free of charge to the user and that is re-usable" (European Commission, 2020a). Scientific information here refers to both peer-reviewed scientific publications and scientific research data.

public policy" (Borgman, 2012, p. 1059) lie in wait. The statement suggests that making data openly accessible will require developments and changes at many levels of the research society. In this thesis, the researcher's experiences are in focus. The accumulation of reliable knowledge is in this context understood as an "inherently social activity" (Borgman, 2007, p. 47), as a fundamentally collective process, a social rather than an individual programme. Seen as a social activity, the communicative aspect of scholarship becomes an important part of researchers' work. To carry out research work involves interaction with a research community, which is why scholarship and communication have been described as "inseparable companions" (Meadows, 1998, p. 159). The web of contexts and relations that constitute the settings for researchers' information activities have been described in several ways. Paisley once drew up a framework of the various systems that all surrounded a scientist, who was put in their centre, pointing out that all these systems had influence on her work (Paisley, 1968). Much has changed since, but the idea that it is not possible to understand researchers' information needs and uses without taking into account the implications of these contexts and relations has been significant for scholarly communication research. Hills later mapped the scholarly communication process as a complex system of information flows with interaction between the scholar (again at the centre of the system), the learned society, the publisher, the product (books, reports, and articles), the librarian and the influence of new communication technologies (Hills, 1983). Each actor had its specific tasks and responsibilities and the interfaces between them were shaped by communication.

Today, researchers' information activities are often described as situated in a larger context of systems or processes referred to as an ecology or ecosystem because of their complexity (e.g. Borgman et al., 2015; Jubb, 2013). A current widely-used term for this ecosystem is *knowledge infrastructure*, which is defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010, p. 17). The infrastructure perspective acknowledges knowledge not as facts and theories found within people's minds or in books, but as "an enduring, widely shared sociotechnical system" (Edwards, 2010, p. 17). On the one hand, technology is needed in order to carry out research; instruments and computers are, for example, used for producing data and emails and journals for sharing the results. On the other hand, the social part of the system is constituted by communities with shared values and norms (libraries, academic departments, theories, support staff) and specialised vocabularies. Included here as parts of this knowledge infrastructure are features like peer-review, bibliographic citation, and evaluation metrics such as journal impact factors. The social and technical aspects of

communication mutually shape each other and are not easily separated. Borgman illustrates the intermingling of technology, practices and social conditions in sociotechnical systems by invoking the example of research data, stating "the tool makes data creation possible, but the ability to imagine what data might be gathered makes the tool possible" (Borgman, 2015, p. 35).

While other studies focus on infrastructural, technological, or economical aspects of the scholarly communication system, this thesis is focused on the social aspects. To succeed in reaching the visions of making research data openly available, research cultures and practices will have to change. Making these changes possible has been identified by many as among most challenging issues that await (Hey et al., 2009; Wallis et al., 2013). Therefore, I direct attention to social and cultural factors, the formation of relations between people and between people and things, and the shaping of understandings and values. The introduction of data policies will potentially change the relationship amongst researchers, respondents, research funders, institutions, and the public. Policies are a form of "epistemic governance" (Jacob & Hellström, 2018, p. 1715). Via the funding instrument and policy, the intention is to change the behaviour of researchers, that is, their data sharing. Studies show a widespread willingness among researchers to share data (Tenopir et al., 2015), but that nevertheless little data sharing actually occurs (Borghi & Van Gulick, 2018; Zenk-Möltgen et al., 2018). External pressures have been shown to erode researchers' control over data sharing conditions and practices, because these can limit the possibilities for making decisions concerning their data sharing themselves (Mauthner & Parry, 2013). Because this phenomenon is recent, researchers' responses to data policies and the implications for how data are shared are research subjects that are currently developing. However, studies have shown that data policies motivated researchers within, for instance, food science and technology and astrophysics, to share their data (Melero & Navarro-Molina, 2020; Zuiderwijk & Spiers, 2019). It has also been observed that engineering researchers who were bound by a data policy did share their data, but without being aware of the data policy's scope and content (Mallasvik & Martins, 2021).

The transition towards open data systems and the new possibilities for moving data and making them accessible are predicted to entail changes in how research data are being viewed and valued. The scholarly publication might not necessarily always be the most important final research outcome. An outcome of research can consist of research data, acknowledged as "significant scholarly contributions in their own right" (Hey et al., 2009, p. 182). However, in contrast to (for instance) research results, data are not easily moved in the first place, as their meaning depends on the apparatuses (software, methods, documentations) surrounding them

(Borgman, 2015, p. 37). More systematic and more controlled data sharing may separate research data from the context in which they were once generated. Even if enhanced with additional metadata and explanatory information, data may risk losing their meaningfulness and in the end become of limited use (Mauthner & Parry, 2013, p. 58). In all this, the researchers' experiences of the nature of data – both what data *is* (as a concept) and what data *are* (as evidence) – becomes another central and important issue that may have implications for how sharing of data, between colleagues within projects or as uploaded in repositories, is carried out.

Aim and research questions

This dissertation situates itself within the area of Scholarly Communication Research of Library and Information Science. The study connects with a branch of investigations that is concerned with issues regarding data management practices, data policy, and open access to research data in the context of the open science movement. The purpose of this thesis is to enrich our understanding of researchers' data sharing in the context of the open data initiatives of external authorities. My aim is to elucidate the researchers' conceptualisations of research data and the ways in which data sharing is an element interwoven in various practices arising from participation in an interdisciplinary community that is bound by a data policy.

In order to achieve the outlined aim, I pose the following specific research questions:

1. Conceptualisations

    How are data negotiated and reproduced within the group? What are data to the researchers, and when?

2. Responses

    How do researchers mutually account for the data policy?

3. Implications

    How do the views on research data and data policies relate to how researchers aim to share data? How can data sharing be imagined as a constitutive yet negotiable element in interdisciplinary research practices?

Research question 1 is directed at the relationship between the researchers and the data in a developing interdisciplinary community. Focus is directed to what data are to the researchers and at what point information becomes data, and how data are conceptualised and collectively made possible within the interconnected relationships. I ask how the concept of data is shaped by negotiation, as a shared practice resulting from participation, and what characterises this conceptualisation. These aspects are explored by studying terminology use, material representations of data, commonly accepted ways of understanding, valuating, and handling data in the various situations of the research process and collaboration.

Research question 2 regards the encounter between researchers and the demands and assumptions expressed in the data policy that they are working with. The question pertains to how the conditions of the data policy enter the researchers' work, which constitutes a meeting of practices that requires negotiations of relationships among those practices. How the community collectively produces a local response to the policy and chooses to shape its meaning is in focus, and activities such as developing the data management plan and understanding the data policy. This research question must be answered in the context of understanding how the researchers define the circumstances of what does or does not matter in relation to data sharing and how this activity is given meaning in daily research work. The circumstances of what to do or what not to do in relation to the data policy, are defined by the researchers' abilities to assess the appropriateness of action within the community.

Research question 3 connects the results from the first two research questions with the data sharing activity, focusing on the consequences that the researchers' understandings of research data and the community's negotiated responses to the conditions of the data policy have for shaping how they take on data sharing. The implications of data conceptions for the participants' data sharing are studied by looking at the researchers' descriptions of daily research practice, such as mutual ways of addressing problems related to what data are to them or to the particular data they work with. The implications of data policy for the community's data sharing are examined by observing whether any change in approach appears to have taken place in the interconnected understandings of what, when, and how data should be shared or not shared. Here, the researchers' explicit expectations of the extent of a data policy's influence become important. I also ask what meanings and intentions appear to have remained stable and continuous after the encounter with the data policy.

By carrying out this investigation, I expect to contribute to Library and Information Science research by deepening and refining our existing empirical knowledge of how researchers collectively view data sharing and give the activity meaning. The new data sharing initiatives have potentially comprehensive implications for how research will be carried out in the future. Increasingly, researchers meet data policies in their profession and must respond to external pressures on data sharing, which motivates the purpose of this thesis: we need to learn more about how researchers' interpretations of and responses to such requirements are shaped. Conceptualisations of data, use of terminology, and identifying when data become data are all examples of important elements included in the activity of sharing data. In order to understand data sharing in research practices, further research is needed that addresses these

intimately related issues. For this aim, the present thesis utilises a Community of practice theoretical lens to learn more about how these aspects are negotiated and collectively shaped within an interdisciplinary research community. Collaborative interdisciplinary research work is currently a strongly promoted way of carrying out research. The findings are expected to be relevant for all partners involved in the knowledge infrastructure that surrounds data management, research support staff such as librarians, policy makers, funders, and researchers.

Disposition of dissertation

In this first chapter so far, I have introduced my research interests and presented the aim for the study, contextualised the subject, and presented the research questions. The remainder of this chapter will introduce relevant related research. Chapter two will present the conceptual framework employed in the study and will introduce the chosen metatheoretical practice theory approach as well as the concept of Community of practice that is central for the subsequent analyses. Chapter three will describe the methodology and methods used for creating and collecting the empirical material. This investigation is a case study focused on one single research group that had to comply with an external authority's requirements on data. I primarily use interviews and observations as methods for data creation. Relevant documents, such as the research group's data management plan, are also part of the empirical material. Chapter four will present the results of the investigation in detail by using a data life cycle illustration to provide structure and support visualisation. The presentation will be divided into three sections, each of which will be followed by an analytical summary foregrounding the most important findings. In chapter five, the results of the empirically based summaries will be brought together for answering the research questions with the help of the theoretical framework and relevant related research. I will end by summarising critical reflections on the results and limitations. Finally, chapter six will provide the conclusions of the thesis. I present potential implications of the results by relating them to the overall context and propose suggestions for future research and field work. An epilogue will end the thesis with thoughts on data sharing from my perspective as a new data collector.

Scholarly information sharing throughout history

To understand contemporary data sharing practices as phenomena, it is fruitful to first look back at history for insights. The structure and praxis of the contemporary ecosystem of scholarly communications reflects, in many aspects, developments and activities from centuries ago. Therefore, a brief history of how scholars have communicated their arguments or findings at certain points of time follows, with focus on a number of the many influences and pressures that have been (and in some cases still are) at work simultaneously, and that shape both the researchers' practices and the knowledge infrastructure system. The presentation aims to contextualise the history of the idea of making research results and data openly accessible, accounting for specific courses of events and developments that occurred during certain time periods.

To share or withhold discoveries – the shaping of decisions

A particularly interesting aspect of the history of scholarly communication that is highly relevant for this study is the variety of influences that, across time, have shaped researchers' decisions of whether to share research information or keep it secret. The economic historian Paul A. David has examined these forces in an essay, aiming to trace the origin of the idea of open science. According to David, an essential, defining feature of modern science is "found in its public, collective character, and its commitment to cooperative inquiry and free sharing of knowledge" (David, 2008, p. 10). These characteristics are followed by an assumption of public knowledge as a natural conceptualisation, an idea that David contests. By presenting how researchers' information sharing developed from ancient Greece to today, dividing history in two eras, David shows that data sharing is a comparatively recent innovation. Without denying that examples of opposing exceptions can be found during each of the two periods, David illustrates that the first era was dominated by what is called an ethos of secrecy and the idea of science as an individual and secretive programme, while the second era was rather ruled by the idea of science as the pursuit of public knowledge. The transition happened during the sixteenth and seventeenth centuries and entailed, according to David, a new set of norms, incentives, and organisational structures.

The ideas that dominated the first era's "imperative of secrecy" (David, 2008) had various origins grounded in competition, morals, religion and tradition. The approach of withholding science can be found as early as the fifth and fourth century B.C. Greece, when philosophers

met in the village Academy outside Athens to declare their arguments. Their ideas and discoveries were communicated via speech and manuscripts. The discussions were written down and then spread widely both within and outside Europe (Meadows, 1998). However, even though scholars shared their thoughts both verbally and in writing, debate was about competition and convincing, which helped shape separate schools of thought, and schools without much interest in sharing simply did not collaborate.

In medieval era, political and religious forces as well as contemporary tradition encouraged that view that science should be considered as pertaining to "the Secrets of Nature" (David, 2008, p. 11) and withheld from the unworthy and vulgar masses. The presupposition of nature as a veiled goddess, Natura, offered moral motives for keeping the information of her matters hidden. Texts about alchemy were regarded particularly important to protect, since the subject was considered divine and potentially dangerous if it should fall into the wrong hands. The solution was to encrypt the information. The small circle of initiated scientists in the early 18th century, including Isaac Newton, used a cryptographic language including symbols and allegories as method to prevent outsiders from understanding their material (David, 2008, p. 15). Even Galileo Galilei was one of many that used the scientific anagram for "communicating" discoveries. Behaving more like a medieval scientist than a second era one (see below), Galileo used an anagram to circulate his discovery of the moons of Saturn in 1610. His anagram "s m a i s m r m i ł m e p o e t a l e v m i b u n e n u g t t a v i r a s" should be read as *Altissimum planetam tergeminum observari* which means "I have seen the uppermost planet triple" (Meadows, 1974, p. 57). It shows how the astronomer managed to establish a priority claim without revealing the discovery itself, and revealing the desire to inform that a discovery had taken place without disclosing what had actually been discovered (Meadows, 1974).

Starting with what David calls the second era, many influences helped form the idea of public knowledge. In the late Renaissance, the reformist movement, in which the philosopher and statesman Francis Bacon played an important role, paved the way for the idea of sharing scientific findings. Around the end of the 16th century, Bacon challenged the rigid philosophical systems and intellectual authority, instead promoting a more collaborative and communicative scientific community (David, 2008). These new thoughts were closely related to the contemporary development of empiricism, another important factor for enabling the idea of knowledge as public. Scholars liberated themselves from the classical scientific systems and started performing experiments. According to David, the interest in exchanging information developed particularly rapidly among mathematicians because of their increased

16

reliance upon new mathematical techniques for solving practical problems in engineering and navigation. As a philosophy of science, empiricism called for open demonstrations of experimental results and evaluation, which encouraged adepts in algebra and experimentalists to develop networks and organise public contests (David, 2008). Communication among mathematicians was intense during the mid-16th century via pamphlets and personal letters. The scholars had started to see the advantages of dividing intellectual labour and drawing upon existing knowledge when solving problems, to see collaboration and information sharing as an asset. This process shaped new cooperative behaviours and norms in relation to the creation of knowledge.

In these times, most scholars were dependant on the patronage of kings, princes, and dukes for being able to work and make a living. Decisions about whether to share or withhold scientific developments like inventions or discoveries were, according to David, rooted in these economic arrangements. The aristocrats generally took one of two positions that were anchored in their motives for tying scholars to the court. The engagement of poets, artists, musicians, and architects at court served both for pleasure and for projecting splendour. Magnificence for instance was manifested through showing off accomplished architects by having them construct grand palaces. David calls these motives "ornamental" (David, 2008, p. 36). Driven by vanity and egoism, and with the will to demonstrate power to competitive peers, these types of creations needed to be widely displayed. Equally important for the ruling families, but for utilitarian purposes, was finding solutions to practical problems. Drawing maps, designing fortifications, and directing water flows in canals were examples of matters that needed engineers and navigators. Because these inventions served partly for protection against enemies, or for triumphing against them in case of conflict, it was of great importance to keep this sort of knowledge secret.

Increased communication among scientists, the development of peer networks, public demonstrations of experiments, and open discussion salons together laid the foundations for an institutionalisation of the more public pursuit of scientific knowledge (David, 2008). State-sponsored academies and scientific organisations were formed in France and Britain as alternatives to private patronage during the mid-17th century, like London's Royal Society, which was formed in 1662 (Meadows, 1998). These organisations created environments where members could exchange knowledge to solve problems. Communication of new ideas was an essential idea for these societies, which launched the first research journals in the 15th

century as a more effective substitute for correspondence and travel.[4] This event is often pointed out as a cornerstone in the development of the modern system of scholarly communication (Borgman, 2015; Jubb, 2013; Meadows, 1998; Prosser, 2013).

This formalisation of scientific pursuit also led to development of regulations of the scholars' communication and norms. Since competition between academies was strong and the issue of claiming priority to a discovery was considered crucial, academies set rules to limit their members' freedom to communicate in order to remain at the top of the academies' hierarchy. Scholars had to ask for approval to exchange information with outsiders. Similar restrictions were introduced in the Royal Society in London, where papers submitted to the secretary Henry Oldenburg had to be read and controlled at Society meetings before possible printing in the Philosophical Transactions (David, 2008). We recognise this process as the system of peer-review used by journals today. These regulations set boundaries between scholars and created a divide between outside and inside the circles, which delimited a group's access to knowledge. Breaking these norms of sharing could lead to grave consequences like destroyed reputations and exclusion.

As time went on, private aristocratic patronage weakened with the growth of these learned societies, the birth of organised learning centres, and the Industrial Revolution. Industry and the state became the new sources of patronage, even as modern scientific research became a university-based activity during the 19th century. However, even in today's settings, similar problems of reputation, agency, and informational asymmetries persist within the academic community (David, 2008).

This brief review shows some of the many factors and conditions that have helped shape the decision of whether to share or withhold scientific discoveries at certain points of history. At times, scholars have had the liberty of taking the decision in their own hands, at times the decision was taken by others. Societal changes, infrastructures, reputational systems, power, norms, competition, priority, religion, and tradition are all forces that have played shifting and contributory roles over time – enabling or hindering sharing – and will seemingly continue to do so onwards.

---

[4] The two first periodical journals collecting scientific European news were both launched in 1665: The Philosophical Transactions of the Royal Society in London, and *le Journal des Sçavans*, by *Académie royale des sciences* in Paris.

Related research

In the following, I connect and position the present study to a larger perspective of scholarly communication research within Library and Information Science that is focused on researchers' perspectives on data and data sharing. The subject of researchers' data sharing is a relatively new research area within LIS. During the years of working on this thesis, interest in these objects of study has increased as manifested in the steadily growing number of published articles, both within the discipline as well as in adjacent fields, for example in Science and Technology Studies, the Philosophy of Science, and Economy and Management Studies.

The selection of studies has in general been guided by the interest of how these two main phenomena in this investigation have been described and studied previously. Because this thesis focuses on the researchers' understandings, experiences, norms and routines, research investigating the social issues related to sharing is emphasized. However, as technological issues including metadata, computing, and knowledge infrastructure all shape the preconditions for enabling data sharing, they each will be touched upon explicitly. The included research was collected during the entire research process and was gradually narrowed down to themes as analysis of the empirical material progressed. The next section is divided into two parts, the first focused on perspectives of the concept of research data, the second on researchers' data sharing activities.

Literature was found by searching in databases (Web of Science, Scopus, Library & Information Science Abstracts) and library catalogues, by browsing scientific journals, via recommendations from peer colleagues, and the snowballing method from reference lists. The search strategy focused on literature within LIS but was open for disciplinary breadth, since studies on data as a concept and data sharing are conducted also in other fields. The literature includes articles, conference papers, dissertations, monographs, and the occasional professional study. Most materials are Scandinavian, British, or American, but literature that originates from or was written in collaboration with researchers from other places, like Germany, Spain, the Netherlands, and South Africa, has been included. The oldest included study of data or data sharing was published in 2003 but most of the materials were published after 2014.

That which is given

Powerful declarations leave little doubt that data are attributed strong positions within scholarly communication research. Data are widely considered to be central for the scholarly enterprise and have been described as "the lifeblood of research" (Borgman, 2012, p. 1066), "a fundamental component in the processes that stabilize science" (Haider & Kjellberg, 2016, p. 160) and "the infrastructure of science" (Tenopir et al., 2011, p. 1). Data have been described as the primary asset researchers have, carrying a value that is to be exploited (Birnholtz & Bietz, 2003, p. 343). It appears that without data, there is no science. Even though it is considered vital for research, there have been surprisingly few attempts to define or explain data that can be found within the scholarly communication literature. This lack has also been noted by Borgman, who commented that data have been generally overlooked in comparison to the many discussions of other scientific phenomena such as facts, representations, or publications (Borgman, 2015, p. 18). In recent years, more researchers, both within Library and Information Science and others, have been directing more interest towards the topic.

The word *data* derives from the Latin etymon *data*, which is the past participle of the verb *dare*, "to give". In such a dictionary definition, data thus means *that which is given* (Oxford English Dictionary, 2018). The term can be used both as a count noun for an item of information and as a mass noun. In scientific texts, the term data was used as early as 1702 as a mass noun for "Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation" (Oxford English Dictionary, 2018). In computing contexts, the term data used as a mass noun has been popular since the middle of the 20th century.

In scientific contexts, the word data has historically been used in plural. Since this use is common tradition among scholarly communication researchers (Borgman, 2015; Haider & Kjellberg, 2016), I will join this convention throughout this thesis as far as the indicated form is evident. The word data will be used in plural when referring to the entities and in singular when referring to the concept.

What are data or when are data?

Within the scholarly communication literature, no agreed or consensus definition of research data exists; it is a concept difficult to outline. According to Borgman, who has dedicated much attention to the subject of the relationship between scholars and research data, one

reason for this is that data can take many forms (Borgman, 2012, p. 1061). Data can be digital or physical, and the sources from which they come or the ways in which they are produced vary widely. In the physical and life sciences, data are mainly generated by experiments, observations, or models, while in the social sciences they are mostly produced or generated by the researchers themselves. In the humanities, data are drawn from records of human culture such as artefacts or archival materials, and the notion of data and use of the term is least developed (e.g. Thoegersen, 2018).

Borgman observed various attempts to define data, of which listing examples of data was the most common way (Borgman, 2015). An operational definition is the most concrete form because it defines data by describing the material called data in a specific operational context. A categorical definition is based on grouping data by various factors, for instance by their origin, value, or degree of processing. In her own studies of data scholarship, Borgman often referred to the United States National Science Board's categorical data definition which is based on origin and preservation value: the observational data, computational data, experimental data, and records (Borgman, 2015, p. 23-24). These categories vary in their preservation value and replicability. Borgman said that while observational data are often the most important to preserve (because they are the least replicable), experimental data may be less important to save (because they can be replicated). However, how data are valued may change in different times, places and contexts, and can be immediate or enduring, gained over time or transient.

In identifying various methods of how to define what data is, Borgman concluded that all methods are arbitrary to some degree, invoking Michael Buckland's phrase "alleged evidence" (Borgman, 2012, p. 1061; Buckland, 1991). In this sense, what data *is* is arbitrary and built upon a common agreement that does not rely on any real proof. To list what *can* be data and what *cannot* is not possible because data are not "natural objects with an essence of their own" (Borgman, 2015, p. 18). Therefore, rather than establishing a moment when something becomes data, data happens in a process in which the researcher recognises that an entity can be used as evidence of phenomena by "a scholarly act" (Borgman, 2012, p. 1061; 2015, p. 18). This act takes place between the researcher and the entity, but is often influenced by external parties that delimit the researchers' space for interpreting data as data.

In line with this constructionist data perspective, Borgman formulated her own working definition for data: "data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship" (Borgman, 2015, p. 28). In this definition, it is the *use* that makes information become data: whoever decides to use a

thing as data also makes data appear. The definition also underlines the aspect of data as representations of phenomena, that is, they are not the evidence in themselves. Borgman says that the interesting question to pose *when* data are, how and when they become data, as opposed to the more commonly asked *what* data are. This question, the author explains, can be studied by observing how individuals or groups create, select or use data and what factors influence these decisions (Borgman, 2015). Borgman's perspective on data is strongly constructionist and will be revisited in the analysis of researchers' notions of data.

One of the few empirical studies that focused on problematising the concept of data was also inspired by the perspective of *when* data are, a question that it investigated by observing the processes around how notions of data emerge and are made possible during the construction phase of new big science facilities (Haider & Kjellberg, 2016). They discovered that the same data – that is, the data of the science facility and its meaning – constantly changed and was relational. The notion of data was shown to depend on various empirical factors, (like the disciplines of the respondents, or institutional function and tasks) and temporality (data passed through many processing stages before they emerged to the researcher as data). Respondents compared data to films that were happening before them – always emerging. The expectations of data and what they could or would contribute to the futures of stakeholders also shaped how data were framed. The authors concluded that data do not exist in and of themselves, but only in relation to other things such as software, instruments, and people. Data are never fixed but "emergent, relational, and shaped by their use" (Haider & Kjellberg, 2016, p. 161).

A contribution written from a Philosophy of Science perspective also offered ideas about what counts as scientific data in analysing what is seen as the paradox characterising the role of data within science. The paper observed that despite their epistemic value as *given* (aligned with the term's etymologic origins), data are *made* (Leonelli, 2015, p. 813). Leonelli suggests also that this tension between viewing data either as human-made or as objective facts can be traced back to the scientific revolution, and that data have neither any truth-value in themselves nor are they straightforward representations of phenomena, which argues for the idea of data as relational. Using the example of annotated genome data (DNA data of organisms) Leonelli wrote that these traces of measurements of samples explain how the idea of raw data was born. Images like these represent the closest we can come to a documentation of a phenomenon in order to use it for informing an inquiry without reproducing or representing the target object itself. Often, the representation has no morphological or conceptual similarity with the original phenomenon and data thus depend on a viewer to

22

interpret them as data and attribute meaning. Data are therefore always "results of complex processes of interaction between researchers and the world" (Leonelli, 2015, p. 813). The results of Haider and Kjellberg (2016) can then be seen as one example of how the meaning of the data change over time during these interaction processes. Leonelli's view also aligns with Borgman's view that it is a "scholarly act" (Borgman, 2012) to recognise data as data a process that includes agreements and interpretations.

In viewing science as a social activity and data as relative to a given inquiry, Leonelli saw data as "tools for communication" (Leonelli, 2015, p. 810-811) whose main function is to enable intellectual and material exchanges. Because these material artefacts must be disseminated, the author suggests that portability is the only essential characteristic tied to data as a precondition for using data. The mundane observation that a physical medium is needed to make data travel has implications because this medium also affects *how* data can travel and *how* data can be used; when media change, the scientific significance may change as well. In stating this, Leonelli is attempting to draw our attention to the physical characteristics of data, because they are seen to have an important role for understanding the epistemic role that data play as social and conceptual functions.

Borgman also discussed the significance of the material representations of data, which all differ in character, even digital data with their various file formats and sizes. Most data practices involve material activities, like handling objects like instruments. Borgman stated that data can rarely be separated from the material forms without also being deprived of the reasons for why they are data (Borgman, 2015, p. 37, 219). In relation to portability and sharing, the author noted that there are sharable and not-sharable forms of data in a broad analysis of data sharing within areas like humanities or science. From a sharing perspective, the data of theoretical researchers have the disadvantage of being difficult to replicate and inconsistent in form and structure. Compared to an experimentalist's machine-collected data, these data can be rather described as structured and consistent but in need of much expertise for enabling interpretation and use (Borgman, 2012). Few other studies have discussed the relationship between the material representation of data and data sharing, and this area has not yet developed as has research into the materiality aspect of information (e.g. Byström & Pharo, 2019; Dourish, 2017).

The perspectives presented here may all relate to a well-known discussion in Library and Information Science, namely the quartet concepts data, information, knowledge, and wisdom, that together form the DIKW pyramid, which places data at the broad bottom, followed by information, then knowledge, and wisdom at the narrow top. The origin of this visualisation in

unclear but it represents one of many ways to display the relationships between the concepts. In his anticipation of a future where data analysis and deductions are automated, Meadows once envisioned that the use of electronic media would lead to a blur between data, information and knowledge (Meadows, 1998). Previously their material forms had separated them from one another; data had traditionally been found in one source (laboratory notebooks), information in a second source (books and journals), and knowledge in a third source (the researcher). As we have seen, drawing lines between data, information, knowledge, and wisdom as well as evidence is indeed not a simple thing to do. The pyramid is rarely explicitly mentioned in current data literature, but provides a reference point for anchoring the discussion.

The results of the presented studies on the data concept provide a useful frame for analysing how the researchers' conceptualisations of data are shaped. Various attempts to frame the concept by using categories, by asking not only what but also when data is, by observing how interaction with data or how data's material representations can have implications and alter their meaning, by studying the value of data as given or made. All these ideas offer different possibilities for identifying how the investigated researchers collectively constituted data.

A multiflavoured concept

There is another relevant empirical study in which the issue of researchers' notions of data played an important role (although this was not the focus of the paper). Chao, Cragin, and Palmer (Chao et al., 2014) made a rare and explorative attempt to map, highlight and display the complexity of research data by looking into the relationships among scientific practices, data concepts, and curational functions. By interviewing researchers within Earth and Life Sciences, they observed disciplinary distinctions in the characterisation and description of terms relevant to data. For example, even if researchers in general frequently use the term *raw data* to indicate data in the early stages of research, this notion varies across fields. To some researchers, raw data meant data coming directly from a sensor; for others, raw data denoted a data set that had had some initial processing; to yet others, raw data was considered data a specific *format*, or a particular *type* of data (in this case DNA sequences). Variations in the notion of raw data were observed not only within the field but even within the same

discipline. For example, some soil ecologists viewed certain measurements carried out when collecting samples as raw data, while others used the term to refer to different measurements.

Based on these collected data, Chao and colleagues developed a vocabulary for specifying relationships among data practices in research, types of data produced and used, curation roles and activities. In this vocabulary, the authors suggest a number of possible categories where features of data are framed, for example *Data characteristics*. This category is further subdivided into *Dimensions* (such as spatial or temporal) and *States* (such as whether data are digital or analogue). Another category, *Data stages*, hosts the subcategories *raw data*, *processed data* or *data represented for publication* (Chao et al., 2014, p. 623). In highlighting these distinctions or flavours of data and forming possible categories and terms for data, the authors move beyond generalities and enrich the understanding of the many contextual factors that influence the use of data. This study's inclusive perspective of the many qualities and characteristics that data can denote, together with the terminology of these data flavours, are useful for the present investigation's analysis of researchers' varying understandings of data.

Data defining boundaries

In an early study of data sharing support, differences in attitudes and approaches toward data between experimentalists and theoretical modellers were observed across several disciplines, and these differences could be used to identify boundaries between groups (Birnholtz & Bietz, 2003, p. 342). Experimentalists collected empirical data and used them for testing theoretical hypotheses about physical phenomena; for their part, the theoretical modellers developed models of physical phenomena in order to compare the models to the empirical data, which entailed that the theoretical modellers needed data as a starting point for developing the models. To the experimentalists, data were the output of an experimental process, while to the modellers, data were input into their work. However, in some disciplines, the same researcher would use data both ways, which means that the differing data use did not always correspond exactly with the membership in a certain group, nor did the particular use of data automatically create a distinction between separate groups' practices. The relationships between the groups were described as "symbiotic" (Birnholtz & Bietz, 2003, p. 342), since they were each dependant on the work of the other. The researchers in both fields negotiated the collaborative relationship to satisfy the needs of both groups.

At the same time, it was observed that there were tendencies among experimentalists to be less willing to share data than the theoretical modellers, who viewed data as more public. The authors concluded that experimentalists wanted maximum compensation from the efforts they had dedicated to generate the data, while the modellers wanted access to the observations that formed the basis of their research (Birnholz & Bietz, 2003).

In an article describing research data sharing as a conundrum, Borgman stated "what data are varies by purpose, approach, instrumentation, community, and many other local and global considerations" (Borgman, 2012, p. 1066). It was claimed that the goal of research (broadly speaking, from empirical to theoretical) is one factor that has implications for what researchers consider to be data. Borgman mentioned that important differences in terminology arose between experimentalists and theoreticians but neglected to further explicate how. However, Borgman also pointed out (as did Birnholz & Bietz, 2003) that the roles are not fixed, because theoreticians might also collect empirical data (Borgman, 2012). In a later monograph, Borgman further elaborated on how documentation of scientific practices within research fields and disciplines develops through long cultural processes: "Data, standards of evidence, forms of representation, and research practices are deeply intertwined" (Borgman, 2015, p. 37). Shared practices and methods may facilitate sharing within a community of practice while it may hinder sharing with people outside.

One of the few studies that asked humanities faculty researchers directly about their conceptions of data also observed that the concept of data differed between researchers from the philosophy department and the others. Philosopher researchers did not consider any of their research materials or what they produced to be data because they described data as "empirical evidence to support a hypothesis" (Thoegersen, 2018, p. 498). Most other researchers however considered some or all of their materials to be data, which was related to their view of data as information related to a question or project. Interestingly, none of the respondents felt certain about the word data. They thought their material did not fit into the idea of data because they compared their material with that of disciplines using numeric and/or quantitative data. Data of a more qualitative character did not fit with how these humanistic researchers thought about the concept.

These studies contribute to the analysis of the results by providing examples of how data practices can differ between groups (experimentalists and theoreticians) that are rarely used as units of analysis.

Data definitions in research data sharing studies

In studies of data sharing, interestingly, it is rarely specified what thing is referred to as data; the data in data sharing appears to be either apparent or implicit. A few authors declare their focus to be on a particular type of data, for instance something general like "raw data" (e.g. Ju & Kim, 2019; Kim & Zhang, 2015) or something very specific like "biological gene expression microarray intensity values" (Piwowar, 2011, p. 2). Other authors focus on the sharing of data collected through specific methods e.g., neuroimaging (Borghi & Van Gulick, 2018). In her study of how ecologists understand and assess data when reusing data collected by others, Zimmerman (2008) used an unusually inclusive definition of data that also included metadata. Zimmerman defined research data as "inscriptions that appear in the form of measurements and observations of the natural world" but also including "information relevant to the data that is independent of the data themselves but without which the data would be incomprehensible" (Zimmerman, 2008, p. 633). This supporting information includes descriptions of methods used to obtain observations or experiments, locations of observations, or attributes of observed species.

These investigations of data sharing offer various perspectives, but what they have in common is a disinclination to explore what form of data is actually being shared and what data are to researchers in the first place, which inspires one to take a closer look at this issue.

What is data sharing?

The activity in focus for this thesis is data sharing. Research data can be shared in several ways. Clarifications of what data sharing means is rarely explicitly explained in data sharing literature; instead, what form of sharing is in focus and what activities are included becomes evident indirectly when results are presented. This observation has been confirmed by a recent meta-evaluation of studies on researchers' data sharing which found that only 21% of the included empirical studies explicitly defined what data sharing is (Thoegersen & Borlund, 2021, p. 5). In some of these studies, the definition of data sharing was quite broad; for instance, one study's definition was "activities involving the dissemination of conclusions drawn from neuroimaging data as well as the sharing of the underlying data itself through a general or discipline-specific repository" (Borghi & Van Gulick, 2018, p. 10). Here, the authors are stating their intention to capture all forms of sharing and all forms of datasets. An even wider definition of data sharing was formulated "any form of release of research data for

27

use by others" (Wallis et al., 2013, p. 2). Others specified data sharing by only declaring which form of data are in focus, for instance raw data (Kim & Zhang, 2015; Piwowar, 2011), even if it may be assumed that raw data are implicitly the main concern of most studies of data sharing studies because they enable reproducibility to a higher degree.

Tenopir and colleagues defined data sharing as that which "occurs when scientists intentionally make their own data available to other people for their use in research or other related scientific endeavors" (Tenopir et al., 2015, p. 3). This definition highlights the purpose of sharing data. The idea is concretised with examples of various activities like including datasets with published articles, posting data on institutional or personal websites, depositing data into repositories, or providing data on request of fellow researchers. To share data as a response to requests from others has also been included in a study of internet researchers' sharing behaviour, where the investigators defined both data set and methods for sharing as when "researchers provide their own data of 'published articles' with other researchers by uploading data sets in data repositories and sharing data sets upon request" (Kim & Nah, 2018, p. 125). To share data after a request from another person is an aspect of data sharing that Meadows classified as "informal" communication, that is, information made available only to a restricted audience, as distinguished from more "formal" ways of sharing, such as publication (Meadows, 1998, p. 7). Different channels for sharing data may have different restrictions and accessibilities. For instance, data repositories may request registration or even a formal request for the data creator to get access to datasets. Access to datasets via journal articles may be restricted by subscription status. The distinction between formal and informal data sharing channels is therefore not easily drawn. However, data sharing via repositories or journals, again, in most cases, does not involve direct contact between data creator and possible user. This form of data sharing between individuals has been called "private sharing of data" (Borgman, 2015, p. 227) and this form of sharing is difficult to document. This method of data sharing will be revisited later on in this chapter.

Prominent researchers of data sharing include the activity under the umbrella of *data activities*. Data sharing is seen as one activity among many related activities, implying a close relation between them; "data activities include the collection, transformation, processing, managing, sharing, preservation and archiving, accessing, and reuse of data" (Palmer & Cragin, 2008, p. 193). Tenopir and colleagues called a similar grouping of activities *data practices* and used an illustration of the Joint Information Systems Committee data life cycle, in which data activities are included with generating and collecting data, managing data, and analysing data (Tenopir et al., 2011, p. 2). Similarly, several studies in various disciplines

investigated data sharing and data reuse simultaneously (Federer et al., 2015; Tenopir et al., 2015; Tenopir et al., 2020; Zuiderwijk & Spiers, 2019) which implies an assumed closeness that connect the activities.

Synonyms for data sharing like *knowledge transfer* (Borgman, 2015, p. 14) and *data release* (Wallis et al., 2013), or the action of *making data mobile* (Borgman, 2015, p. 219), have been proposed. In studies of the social aspects of making data accessible, these have appeared only rarely.

Factors shaping the decision to share data

Many studies have shown that researchers are generally willing to share data when asked. For instance, an empirical study conducted over four years (2009–2013) showed an increased general willingness and positive attitude towards sharing data among international researchers in 19 disciplines (Tenopir et al., 2015, p. 7). The will to share some or all their data in repositories without any form of restrictions had also increased significantly, as had agreement with the idea of creating new datasets from shared data and using data in ways other than intended. Not only the researchers' attitudes appeared to have changed, but also the actual data sharing practices. Nearly three-quarters of the respondents stated they made some of their data available to others in the follow-up, which was a significant increase compared to four years earlier (Tenopir et al., 2015, p. 8). Data were thus increasingly being shared. A later survey of geophysicists' data sharing confirmed the positive attitudes to data sharing; 82% of the respondents stated they would be willing to place some data in a central repository (Tenopir et al., 2018, p. 896).

Despite demonstrations that attitudes have implications for actual behaviour (Kim & Burns, 2015; Kim & Zhang, 2015), it appears the general willingness to share is not reflected in practice, and many studies have suggested that not much data sharing appears to actually be carried out (Borghi & Van Gulick, 2018; Borgman, 2012; Darch & Knox, 2017; Fecher et al., 2015; Tenopir et al., 2011; Wallis et al., 2013; Zenk-Möltgen et al., 2018). Apparently there is a discrepancy between ideal and actual practice despite the introduction of data policies. Borgman concluded that "willingness does not equal action" (Borgman, 2015, p. 205) and warned of the risk of drawing conclusions only from researchers' statements of their attitude and sharing practices.

Attempts to measure the extent of data that are being shared have looked at data sharing carried out via scientific journals. Results showed that only 13% of published articles in the top ranked journals in biology, chemistry, mathematics, and physics made at least some of the original data available (Womack, 2015, p. 19). Another study found that around half of the empirical papers in sociology and political science journals stated that the data were available, but when these statements were followed up by clicking on the links, only 37% of these data sets could actually be accessed (Zenk-Möltgen et al., 2018, p. 1053). The authors of these two studies concluded that there was a progress of sharing data in journals but that it was slow despite journals adopting data policies. As for data sharing in repositories, 60% of the researchers within the biomedical research community stated they had never uploaded data to a repository (Federer et al., 2015, p. 8).

An array of factors has been shown to have implications for researchers' decisions to share data and sharing behaviour, for instance, perceived motivators, barriers, and risks. Among the factors motivating researchers to share data is the will to contribute to scientific progress (Linek et al., 2017) and foster research transparency and enable reproducibility (Borghi & Van Gulick, 2018; Melero & Navarro-Molina, 2020), the desire to help other research colleagues (Thoegersen, 2018), and advancing a specific area of science (Federer et al., 2015, p. 9). The most highly ranked motivators vary between studies and disciplines. The availability of repositories was found to positively influence internet and STEM (science, technology, engineering, and mathematics) researchers' actual data sharing behaviours (Kim & Nah, 2018; Kim & Zhang, 2015). Previous experience of sharing data has proven to be another factor that influences data sharing, and researchers within sociology and political science who had shared data in journals previously tended to repeat this activity (Zenk-Möltgen et al., 2018). Results have further shown that authors publishing in open access journals were more likely to share data (Piwowar, 2011).

One factor that is often indicated as hindering researchers from sharing data across different disciplines is lack of time (Borghi & Van Gulick, 2018; Tenopir et al., 2015), and one study even named this factor to be as the greatest hinderance for data sharing (Tenopir et al., 2011, p. 9). The labour involved in making data interpretable, for instance by postprocessing them, makes researchers less likely to share their data (Borgman, 2012, p. 1066), and making data re-usable by others is perceived to require much effort (Melero & Navarro-Molina, 2020, p. 7). These results can be related to the observed fear that geophysicist researchers had that their data could be misinterpreted or misused if shared (Tenopir et al., 2018). Difficulties in choosing a suitable repository (Federer et al., 2015) and

lacking knowledge of existing sharing infrastructure (Zenk-Möltgen et al., 2018) were other impediments that kept researchers in biomedicine, sociology, and political science from sharing data.

Issues related to generated data itself have also been shown to form obstacles for data sharing. The perception that people did not need their data has been an issue which restrained researchers across disciplines from sharing data (Tenopir et al., 2015). In astrophysics, not knowing whether the data were useful to people was found to be an important factor for the decision to share data, and the perception that the data were useless made researchers consider sharing data openly a waste of time (Zuiderwijk & Spiers, 2019). At the same time, the same study found that knowing that the data were useful to others was among the most important motivators for data sharing.

Other data-related issues that form barriers for data sharing are confidentiality and sensitivity. Researchers within neuroimaging were concerned that they had no right to make data public, as these data contained confidential or sensitive information (Borghi & Van Gulick, 2018). Borgman reflected that ownership and rights indeed constrain the availability of data, but at the same time, researchers often treat data as intellectual property whether they are or not (Borgman, 2015). Researchers in the humanities showed little or no concern about intellectual property, privacy, or security issues in relation to the material they used. Instead, their anxiety concerned issues related to confidentiality and the topic's sensitivity (Thoegersen, 2018). Among engineering researchers, having industrial partnerships has been demonstrated to deter researchers from openly sharing data. This form of collaborative work often entailed that generated data constituted commercial secrets which were bound by carefully monitored non-disclosure agreements (Mallasvik & Martins, 2021). One study found that the most distinct pattern related to subject discipline was that researchers who worked with human subjects data (in Medicine and Health Sciences) were much less inclined to share all or some of their data without restrictions in an open repository than those of other disciplines (Tenopir et al., 2015). Data containing sensitive information was in the same way the main reason for not sharing data outside of the one's own research project in a study across four domains (arts and humanities, social sciences, medical sciences, and basic sciences) (Akers & Doty, 2013).

Competitive factors and award structures are other obstacles that have been shown to shape researchers' data sharing. The fear of making data openly accessible when they potentially contain additional findings to discover has been observed by several studies (Borghi & Van Gulick, 2018; Fecher et al., 2015; Thoegersen, 2018). Astrophysics researchers stated that the

high value the academic system puts on publications rather than datasets made publishing a priority (Zuiderwijk & Spiers, 2019). The need to publish before making data available was the new perceived top-ranked barrier that replaced the former main concerns like lack of time or funding (Tenopir et al., 2015). The need for publications was expressed particularly by biologists and physical scientists, while for computer scientists and researchers in Education and Information Science, this concern was a lower barrier to sharing data. Borgman found that policies have failed to acknowledge factors like the competitive nature of scholarship, or the importance of incentives for reward (like citations) when promoting openness (Borgman, 2015). A number of studies across different disciplines have shown that an increased likelihood of article citations or formal acknowledgements for shared data is an essential motivator for getting researchers to share (Borgman, 2012; Chawinga & Zinn, 2019; Melero & Navarro-Molina, 2020; Tenopir et al., 2015; Zuiderwijk & Spiers, 2019).

Finally, preferences of data sharing methods have been observed to affect the decision to share or not. Several studies have shown that sharing data informally (the "private" data sharing mentioned above) as a response to requests from other researchers is a common form of sharing (e.g. Akers & Doty, 2013). In another study, the conclusion drawn was that exchanging data personally was common and made the data valuable because documentation and tacit knowledge was gained (Wallis et al., 2013). A condition for sharing data in this study was however that the requestor was known and trusted. These factors increased willingness and efforts to prepare data because this data sharing was viewed as a "peer-to-peer relationship" (Wallis et al., 2013, p. 13). Similarly, researchers often consider sharing on request to be like assisting a known colleague or as collaboration (Federer et al., 2015). Borgman found that sharing data with known persons, and expecting immediate use, was easier than making data accessible for unknown persons and for unknown times (Borgman, 2015). In addition, distance from others – geographical, temporal, methodological, and regarding levels of theory and expertise – matter, because these increase the difficulties of interpret data, which might be one reason favouring private data sharing. The closer a researcher's work is to the origins of the data along these dimensions, the less the interpretation depends on formal knowledge representations (Borgman, 2015)

This collection of influences on whether researchers share data are valuable for getting a broad picture of what is known about researchers' relationships to data and data sharing. Knowing about motivations for and barriers to sharing data is relevant for analysing the investigated researchers' approaches to this activity.

Data sharing practices and data policies

Policies regarding data sharing are currently being adopted by stakeholders like scientific journals and research funders, and in response, research about the affect these policies have on research practices is developing. Most studies discuss the influence of data policies as one among others that influence researchers' data sharing practices but without focusing on this relationship specifically. Because the current moment is a transitional period when data policies are appearing as a new preconditions for researchers, investigators looking into researchers' attitudes and perceptions need to account for whether a data sharing policy already existed or not when the observations of the researchers took place (e.g. Tenopir et al., 2015).

A number of studies have shown that policy with requirements for data sharing is a factor that shapes researchers' data practices, even if no study indicates that a policy was the most important motivating factor. For instance, researchers in astrophysics have said that introducing mandating policies would make them share their data (Zuiderwijk & Spiers, 2019). Within food science and technology, compliance with the policy of funders was observed to be an important reason for deciding to share data (Melero & Navarro-Molina, 2020). In neuroscience and biomedicine, policies were observed to be weaker motivators than other factors (Borghi & Van Gulick, 2018; Federer et al., 2015). An investigation that focused specifically on the relationship between sociology journals' data policies and researchers' data sharing practices found a positive correlation between present data policy and data being made available and accessible along with the published results (Zenk-Möltgen & Lepthien, 2014). In empirical papers in journals with a data policy, as many as 75% of the authors stated that data were made available. However, results also showed that as many as 56% of authors shared data even in journals without a policy. Since expectations about availability of data for replication analyses were far from fulfilled, Zenk-Möltgen and Lepthien concluded that more specific and detailed policies are needed that more clearly explain to researchers which data are expected to be shared and how.

In a later study, the same group found that similar numbers of journal articles stated their data had been made available, but in addition, the actual availability of the data was checked. The results showed that only 37% of the data of the empirical articles could actually be accessed (Zenk-Möltgen et al., 2018, p. 1059). Again, a gap between the requirements of data policies and reality came into light. The authors neglected to further explain these findings, but concluded, in line with their former study, that journal data policies can play a major role

33

in closing the gap between policy and practice via establishing solution-oriented policies with clear criteria and implications (Zenk-Möltgen et al., 2018).

A few years earlier, a similar study of data sharing in 40 high-impact journals in biology, chemistry, mathematics, and physics had also shown a low degree of data sharing in articles and, in line with Zenk-Möltgen's group, found that the availability of data had not grown much despite the presence of or phasing in of data policies (Womack, 2015). The results also showed that separate DOIs (Digital Object Identifiers) and direct links to data were rarely used in the analysed journals. More typical was the use of a loose style of linking to data without complete references and few details about the data. Further, few articles proved to link to original raw data as originally envisioned by data sharing advocates (Womack, 2015). The issue of to what degree data policies are lived up to is still rarely an object of study, but a systematic review of data sharing globally observed that researchers not following policies were generally not reprimanded (Chawinga & Zinn, 2019). The explanation given was that funding organisations are inconsistent in enforcing their own policies and ensuring contractual obligations.

Another study attempted to evaluate researchers' data sharing practices by following project data via the data management plan, and found that sharing was not carried out in the majority of cases (Van Tuyl & Whitmire, 2016). When data were shared, access, documentation, or formatting were often questionable. The authors concluded that funding agencies need to set minimum definitions and expectations for sharing. A very recent investigation of engineering researchers in UK and Norway showed that policy and funding requirements affected how and when data sharing occurred (Mallasvik & Martins, 2021). However, even when they were bound by data policy, researchers were unaware of the content and scope of these policies. Complying with a policy was thus carried out instrumentally. An explanation the researchers gave for this was the instructive nature of the requirements.

Instead of attributing the reluctance or resistance of researchers to share data to obstacles in methodological, ethical, legal, scientific, or technological infrastructures, a study in Management Studies focused on the data policies as such. The authors questioned the view of data policies as neutral, passive, and innocent mechanisms, and analysed their underlying scientific, moral, and political frameworks. The authors argued that policies evidently embody and institutionalise a particular understanding of scientific and data sharing practices, norms and ideals; most policies focus on requiring full and open access to data. This internalised perspective was thought to have implications for researchers' work, which according to the

authors creates a tension with understandings of data and data sharing. A more encouraging "case-by-case approach" (Mauthner & Parry, 2013, p. 62) to archiving and sharing data was offered as alternative; this would facilitate for researchers and leave space to make tailored choices for specific data and studies themselves.

Formulating data management plans is usually included as a requirement of data policies. A recent investigation showed that researchers, policymakers, service providers and research support staff involved in data sharing had different views on these plans, which varied according to the researcher's background or work situation (Kvale & Pharo, 2021). Researchers working in collaborative environments where data sharing between colleagues or with external partners was a common activity used the data management plan as an internal document for how to agree on standard procedures. As a standardised form, the data management plan helped facilitate translation among various stakeholders "by creating context for research data so that these could be understood and interpreted in the different worlds" (Kvale & Pharo, 2021, p. 15) across disciplines, institutions or national boundaries In contrast, researchers who worked individually or in collaborations that involved little data sharing, were more focused on data documentation and personal privacy aspects.

Reflecting on why perceptions of sharing and sharing practices among researchers in certain areas remained unchanged after four years, Tenopir and colleagues assumed that it might simply be that practices take years to change, or that the effectiveness and impact of data policies, plans and infrastructure has to be questioned (Tenopir et al., 2015). They speculated about whether a tipping point might have been reached where attitudes remain the same despite pressures implemented. Additionally, existing data practices related to promotion and collaborations might be too dominant to enable a change of attitudes and practices.

Research discussing the implications that policies concerning data sharing have on the work of researchers is highly relevant for this thesis. The gaps between policies and practices further motivate the importance of in-depth perspectives of researchers' understandings underlying their data sharing practices and perceptions of data policies.

Chapter summary

To highlight the most important findings which will be used in the thesis, I end the chapter with a conclusion of the studies. The overview of how previous research has addressed the

data concept from various angles showed accounts of possible ways of categorising and defining the many existing forms data, shaped by the many existing ways for producing them. Attempts have been made to categorise data features, such as their characteristics and stages. Conceptualisations of the term *data* have been shown to depend on research method and discipline, and even to vary within the same discipline. What data *is* has been observed to be emerging and related to the concept of change along with the function that the data had, the work task within which the data were used, and the expectations about the role the data would play. Physical characteristics have been highlighted that have been shown to have implications for how data are viewed, used, and communicated. Furthermore, it appears that the concept of data is frequently being formulated from either of two different perspectives. In the former, data become data in an act of recognition between researcher and data; in the latter, data are rather seen as given and having inherent values. The public widespread positivist view of data as given, quantitative, and numeric has been observed to make researchers who use qualitative data consider their material not fit well with the concept of data.

Regarding the data sharing activity, studies have shown that while many researchers agree with the idea that sharing data on a larger scale may be important and beneficial, and while it is observed that data are increasingly being shared, the development is generally thought of as being slow. Several studies have investigated factors that have implications for researchers' data sharing in terms of motivators or barriers. Competitive factors, data sensitivity or confidentiality, data usefulness, method of sharing, time and effort needed for sharing data, and availability of repositories, are all some of the issues that have been observed to shape researchers' data sharing. It is noteworthy that few empirical studies of data sharing clarify either their definition of data or method of sharing referred to. Investigations of the effects of the data policies that are currently being adopted have shown that the presence of data policies motivate data sharing differently in varying disciplines. The data policies of journals have been observed to have little effect on the amount of data that are being shared, and authors have called for clearer policies that explain for instance what data are expected to be shared. The uses of and views about the data management plans frequently included in the data policies' requirements have been found to vary depending on whether researchers work in collaborative or more individualised environments. The slow changes shown in researchers' data sharing activities, despite implemented data policies, have raised questions about their effectiveness.

2. Conceptual framework

In this thesis, research work generally, and the shaping of data and data sharing specifically, are addressed as social activities. To carry out research in most cases involves interaction with one or more parts that constitute the larger research community. The theoretical concept *practice* is central for guiding this case study's research design, data collection, and analysis; a practice approach thus forms the theoretical foundation. As an instrument for actively analysing the empirical material, I employ the *Community of practice* concept. These concepts form the theoretical lenses through which I approach the study object, and they will now be described in more detail.

Practice theory and knowing

Practice theories is a theoretical perspective currently applied broadly in Library and Information Science and several cognate disciplines. There are a variety of different takes on practice, both methodologically and theoretically (Pilerot et al., 2017). The approach has been used for example to study professionals' knowledge and learning (Pilerot & Lindberg, 2018), citation behaviours (Gullbekk & Byström, 2019) and enactments of environmental issues (Haider, 2011). Around the 1950s and 1960s, studies on scholarly communication gradually started to shift focus from system-centred studies toward more person-centred studies (Talja & Hartel, 2007); rather than analysing researchers' end products, interest was directed at observations of the ongoing research processes in their contexts (Palmer & Cragin, 2008, 169). A practice approach was applied early within the field of Science and Technology Studies to study science and researchers' work processes in the environments where this work happened, in laboratories and institutions, by using case studies and ethnographic methods. In Library and Information Science, the interest in studying scholarly activities where these were locally situated developed more slowly, and survey research remained the dominant method for studying information practices for a longer time.

Practice theories, or the *practice theory approach*, is not a single or unified theory, but is rather represented by a "broad church of theories" (Cox, 2012b, p. 63) comprising many varying definitions and debates. Although no single theoretician can be referred to as the originator of practice theories, Karl Marx and Ludwig Wittgenstein have been pointed out as major cultural roots whose ideas have made significant contributions to the approach (Nicolini et al., 2003). Unifying for the different branches within the family of practice

theories, and central for this thesis, is the idea of organisation of activities and the notion of knowledge, and the connection between them.

The philosopher Theodore Schatzki laid the foundation of practice theory in the form that many investigators choose to use today. Schatzki views social life as shaped by activities organised or interrelated in a certain way, and the basic units of activity are defined as "bodily doings and sayings" (Schatzki, 2002, p. 72). Things like movements of limbs and the meaning of things people say are thought of as gathered into "bundles" (Schatzki, 2002, p. 71) of activities. What in turn merges these bundles of activities and unites them into practices is described as a "temporally evolving, open-ended set of doings and sayings linked by practical understandings, rules, teleoaffective structure, and general understandings" (Schatzki, 2002, p. 87). This definition has been translated as including knowing how to do an activity, the activity's purpose, and explicit beliefs about how activities are to be carried out. For instance, the definition can include instructions (Cox, 2012b) or anything about an activity that is said in terms of "one must… one should… one usually…" (Pilerot & Lindberg, 2018, p. 256). Thus, knowledge, norms, and conventions are what unite activities into practices; this form the premise for how the practice approach is used in this context. More specifically, knowledge is seen as "situated in the system of ongoing practices of action, as relational, mediated by artifacts, and always rooted in a context of interaction" (Nicolini et al., 2003, p. 3). This perspective of knowledge puts the social in focus; learning is viewed as a social rather than a cognitive activity. The belief of the value of persons' doings stems from Marx's epistemological principle of human activity and the world as interconnected, that knowing cannot be separated from doing and that learning is a social and not only a cognitive activity (Nicolini et al., 2003). In being an ongoing performed practice, knowledge is, through the practice theory lens, conceptualised as a verb rather than a noun, as *knowing*.

How practice theory perspective explains social life and views knowledge aligns with the aim of this thesis to elucidate how researchers understand and shape research data sharing collectively. This view entails that researchers' sayings and doings are the main unit of analysis, allowing observations of data and data practices from a researcher perspective, and entering the researcher's world and understanding(s) of data and data sharing. Researchers' practices express understandings that are not identified as either subjective or objective and which are interpreted and explained in relation to contextual factors.

The choice to employ a practice approach as metatheoretical standpoint was made at the planning stage of the investigation. As an active take on practice and concrete instrument for analysing the empirical material, the Community of practice concept is utilised. This concept

focuses on articulating a connection between group and practice, guiding attention to dimensions like meaning production, and explaining significant phenomena (Wenger, 1998).

Community of practice theory

This study employs a *Community of practice* perspective as theoretical tool for analysing and explaining the empirical material, the most well-known of the practice theories within Library and Information Science (Cox, 2012a). The initial concept was originally formulated in collaboration with Jean Lave (Lave & Wenger, 1991) and further developed into a theory of practice by the educational theorist Etienne Wenger (Wenger, 1998). The theory models and operationalises learning by relating it to social structures and to participants' practices in the processes. Community of practice is one of the most influential conceptions of social learning to date, and has been widely applied within Library and Information Science and related disciplines like education and organisational studies. The idea has for instance been used to design inclusive development programs for professionals in academic libraries (Carroll & Mallon, 2021), to investigate learning processes among farmer groups converting to organic farming (Morgan, 2011) and to study social norms around data generation and consumption within science and engineering (Birnholtz & Bietz, 2003).

Wenger defines *practice* as "doing in a historical and social context that gives structure and meaning to what we do" (Wenger, 1998, p. 47) and the concept is thus coherent with a general practice approach. Practice includes the explicit, such as language, documents, defined roles, and procedures, as well as the tacit, such as relations, the unsaid, perspectives, understandings, and assumptions. Shared practices are key components to a community of practice because they, according to Wenger, shape "the source of coherence" (Wenger, 1998, p. 49). This relationship — the connection between community and practices — defines the community and is viewed as associated by three dimensions: mutual engagement, joint enterprise and shared repertoire (Wenger, 1998, p. 73). Each dimension has its own characteristics, but all three are interrelated. In this thesis, I will employ the dimensions which form coherence to identify shared data practices within the research group which is a frequently used way of using this theoretical framework. For instance, a review of studies in online learning research using this theory showed that a majority applied the dimensions to investigate if and to what extent community of practice had formed (Smith et al., 2017). Within LIS, the same approach has been used to analyse the dynamics of information practices; for

example, this approach was used in an investigation into shared ways and understandings that were shaped in the setting of a Facebook group (Mansour, 2020).

First, mutual engagement defines a community, because engagement is what forms relationships among people when they unite in shared activities (Wenger, 1998). Members of a community shape acceptable ways of having relations through negotiation, which enables the community to cohere and function. The social relationships form an identity because the participants share and experience competence that distinguish them from people outside of the community. In later formulations of the theory, competence is defined as describing "knowing negotiated and defined within a single community of practice" (Wenger-Trayner & Wenger-Trayner, 2014, p. 13). Competence of members within the community may be "complementary" or "overlapping" (Wenger, 1998, p. 74) but have to be valued and recognised by other members, as does each member's specific contribution.

Because members of a community of practice are connected by engagement, relationships among them can be complex. To be a member of a community of practice means, according to Wenger, to be committed to the work needed to sustain relations that are organised around the common activities and make mutual engagement possible. Mutual relationships may involve tensions and disagreements. Therefore, efforts to link the boundaries of different practices, for instance "brokering" (Wenger, 1998, p. 109), nurturing and coordination of activities, are needed to enable the community of practice to cohere and function over time.

The second dimension addresses the common ways of engaging in activities in a community and how these ways are negotiated and agreed upon, also known as the joint enterprise. Within a community of practice, the enterprise "directs social energy" (Wenger, 1998, p. 82) by forming common grounds. This mutual accountability defines the circumstances of how things should be carried out, how goals should be reached, what is important, and what does or does not have value and make sense. An enterprise may be explicit in formulated goals and rules, or implicit in perceptions or in the ability to assess appropriate action or qualities. The participants define the enterprise through their mutual engagement in practice. Each member's response results from an individual choice. However, because all members are interconnected, the "relations of accountability" (Wenger, 1998, p. 81) are formed as a collective response or interpretation of the conditions created by the contexts within which the community develops.

Wenger states that while a community negotiates its own mutual accountability, the community's practices are also formed and conditioned by the world surrounding it, "the broader system" (Wenger, 1998, p. 79). Social, cultural, political, and institutional conditions,

40

for instance, can lie beyond the community's control. However, because Wenger's idea of practice is that it is actively produced by participation and agency, the community members have space to produce the response to or interpretations of outside pressures. In this perspective, no single condition can thereby directly determine the community's practices.

The third dimension of coherence is the shared repertoire which represents the resources that are produced over time within a community engaged in a joint enterprise. These resources are united by how they reflect the community's history of mutual engagement and relationships; they are part of practice and belong to the community as significant components of daily work. Examples of a community's repertoire are experiences, routines, use of terms, tools, conscious or unconscious ways of doing things and addressing problems, and concepts that reflect specific perspectives of the world (Wenger, 1998). That the resources belong to the community and are recognised by members does not mean they need to be locally created; Wenger points out they are often imported, adopted and adapted (Wenger, 1998). Like the joint enterprise, the meaning of the resources is not seen as static, but as continuously and actively negotiated by the participants (Wenger, 1998).

Strengths and limits of the theoretical concept

The widespread use of Wenger's community of practice concept in several fields has demonstrated the theory's many strengths. Since it was introduced, the author has continued to develop the concept, and later texts formulate different specific definitions of what a community of practice is. For example, a more recent definition is that a community of practice is "a learning partnership among people who find it useful to learn from and with each other about a particular domain" (Wenger et al., 2011, p. 9). The three dimensions connecting practice and community continue to be present in Wenger's later descriptions of the theory and these have taken, in my view, even more concrete and applicable forms. The connections have been described as three crucial characteristics of a community of practice: *domain*, *community* and *practice* (Wenger-Trayner & Wenger-Trayner, 2015; Wenger, 2004). The empirical content of the dimensions appears to be relatively unchanged.

As a theory that holds a strong position and is broadly cited and utilised, Wenger's idea has of course been much debated, and certain aspects of the concept's limitations have been critiqued. To begin with, the use of the term *community* has been controversial and criticized by many as problematic because the word connotes harmony and consensus. For instance,

Cox has argued that Wenger's conceptualisation of what a community is, is "paradoxical" (Cox, 2005, p. 532) in that Wenger uses the concept to invoke connotations that goes against many general assumptions about the term; a community is not necessarily friendly or homogenic, nor agreeing and safe, nor a social category or unit, not without purpose, and not static but creative. In fact, this list of issues addresses and summarises many aspects of critiques raised against the theory.

One line of criticism has been directed towards the idea of consensus in a community of practice, that is, that participants are seen to agree and share perspectives. This issue is one that Wenger addressed when formulating the theory. Even though the community was to be seen as a "home base of the practice" (Wenger, 1998, p. 123), this did not necessarily entail community homogeneity or that all practices were thereby shared. According to Wenger, not everything the participants do has to be accountable to a joint enterprise, nor do all participants have to interact with everyone or be geographically close. However, the less that each of the three dimensions is present, the less it is probable that a community of practice has actually formed (Wenger, 1998).

In Cox's view, the strong focus on what is shared within a community makes the theory tend to overlook the richness and variety of thinking that can be found in many communities. However, to direct attention to the "common ground and perspective" (Cox, 2012a, p. 180) that may exist among persons who engage in a practice, is one of the theory's primary strengths. In relation to the debate of the theory's strong emphasis on community, several researchers have proposed reversing the concept into *practices of communities* in attempt to put the activities in focus rather than the group. By such a shift, the subject is decentred and instead becomes the result of an interplay among several actors, "the material world, knowledge, and the actors present" (Gherardi, 2009, p. 528).

Another controversial and much debated aspect of the community of practice theory is how the theory neglects to include power as an element, which takes us back to the idea that a community of practice is an agreeing unit of people. Seen as a unit that negotiates its own rules and understandings, a community of practice has been interpreted as if existing in isolation from the context that surrounds it. In summarising themes or critiques aimed at the theory, Gherardi acknowledges that it does not consider power relations, neither "within organizations or between these and external sources of knowledge/power" (Gherardi, 2009, p. 521). To the critique of the concept as powerless, Wenger has responded that power issues are inherent to and at the very core of the perspective, even though the theory is indeed not a political one but is about learning: "every learning move is a claim to competence" (Wenger,

2010, p. 186). Wenger also discusses these power dynamics that operate both within and outside of a community of practice by using terms like "economy of meaning" (Wenger, 2010, p. 186).

Regarding power and how communities of practices exist in relation to each other, in the original monograph, Wenger presented the perspective on context around communities of practice as a "social landscape" (Wenger, 1998, p. 118). This landscape is defined by the boundaries of various practices, boundaries that become articulated and apparent when meeting. Situating a community within a landscape allows investigations into what happens when practices encounter each other and of the varying connections forming between them, which is another aspect highlighted as one of the theory's strengths (Cox, 2012a). Because different practices have differing abilities to influence the landscape, the landscape of practices is in later texts described as evidently political and hierarchical. One practice may have the ability to influence another practice through legitimacy of discourse or control over resources. National policies are given as an example of "attempts to colonize the field of practice" (Wenger-Trayner & Wenger-Trayner, 2014, p. 15-16). The power position of one practice leads to governing the practice of another.

The question of the amount of agency enjoyed by community participants is another subject for discussion. The theoretical perspective has a strong focus on the participants as creating meaning through active participation rather than by passive reproduction, and thus as able to negotiate their responses to both internal and external claims of competence. According to this view, issues like how understandings may be formed by routines or by the work task themselves have been ignored (Cox, 2005). For instance, Cox points out that contemporary work circumstances such as frequent reorganisation, strong management, and individualised work limit the possibility of actively forming local practices. These circumstances, Cox suggests, entail that community of practice is best applied when analysing problem-solving situations where a local practice is allowed or expected to be created, rather than on situations of routine work. Wenger has responded to Cox's suggestions that the concept is anachronistic by writing that the concept is not meant only for application to specific moments of organisational work; however, Wenger admits that the notion of community of practice will manifest differently as the society evolves (Wenger, 2010).

Operationalising theory

In this thesis, the investigated group of researchers is viewed as a community of practice and focus is on the potential data sharing within the researchers' activities. A detailed description of this project group will subsequently be presented in chapter three. Directing attention to the common experience neither entails that all researchers' views are shared nor that the individuals are ignored. Each researcher evidently forms her/his own response to what s/he experiences as being the situation, but because the participants are interconnected, they are each other's concern and have to find ways to coexist and agree.

The investigation's focus is on how data conceptualisations and data sharing are activities interwoven as elements in the group's social activities. In such a perspective, data are seen as playing a role in the researchers' social practices rather than being the goal of these practices (Cox, 2012a, p. 186). To explore how researchers conceptualise data and collectively make data possible, how shared views of data exist and have been formed, I see the data concept as a resource and thus as part of their repertoire. This approach allows me to study the researchers' specific ways of addressing data as concept, whether conscious or unconscious, as well as their use of the term "data." In addition, characteristic features of the project data, such as their material representations of data, will come into light. These perspectives thus reflect the researchers' specific understandings of the concept, understandings that can either be imported or created within the community itself. Viewing data conceptualisations as belonging to practices means these may highlight boundaries between different communities of practice. The creation of boundaries will be actualised in the analysis of the data as resources, as will the issue of multimembership, that is, how researchers belong to many communities of practice in parallel. Data activities thus belong to these identities and may articulate a disconnection between internal communities of practice in the form of lacking routines and understandings. The dimensions of mutual engagement and joint enterprise are useful for examining if and how researchers work actively to form an environment where data as a concept is made possible, or negotiate their mutual account of what data is.

The encounter between the data policy and the research group is here interpreted as a situation where the contextual conditions of "the broader system" (Wenger, 1998, p. 79) that surround the community attempt to influence the group's practices. The policy was launched by the funding agency, which has the ability to make claims to competence because it controls the resources that finance the research project. The relationship between funder and project group is thereby seen as hierarchic and involving power dynamics; the funder has the power

to challenge the research group to renegotiate and change its practices. The research group in turn has to respond to these imposed external claims. When the community of practice needs to deal with external demands, a problem-solving situation—an occasion for learning—is created. Because I view the researchers as active participants who always have the power to negotiate a response as a local practice, the encounter with the policy is not a question of simple implementation. The funder cannot directly influence the researchers' practices. Rather, the researchers have to negotiate by deciding which aspects of the data policy they accept and which they refuse, and which changes to existing practices are seen as meaningful and which not.

By identifying what is valuable and makes sense to the researchers, and observing how they evaluate the appropriate actions to take in order to reach their goal of shaping a response to the policy, the dimension of joint enterprise becomes actualised. Particularly important is the concept of mutual accountability, which includes shared accepted 'ways of doing' that define what data to share or not, when to share, and how, within this particular setting. Mutual accountability forms the basis for how the community chooses to deal with the external pressures that challenge their existing practices. Their mutual engagement allows me to explore the efforts researchers have to make to shape an environment where it is possible to develop an interpretation of the data policy, and also explore how different competence are recognised and made use of. By viewing the data management plan (DMP) and other documents the researchers produce when dealing with the data policy as resources, I can identify agreed perspectives of what is considered meaningful; what is prioritised and what is not. The DMP will show how data management preserves aspects of the community's tacit knowledge and helps initiate newcomers into existing practice as they gain access to those practices. It will become evident that all dimensions of coherence are interlaced and have implications for each other.

3. Methods and materials

This chapter presents the overarching methodological approach and the methods chosen for data creation and collection. The empirical material for the study consists primarily of transcribed interviews, observation notes, and a specific document.

The case study as research design

This study is a case study, which is one of the most commonly used approaches when conducting qualitative research (Merriam & Tisdell, 2015). While it sounds straightforward, it is anything but. Even if it is often used, it is also one that is much debated and comes in many varieties, to say the least. Case studies have been used, for example, to take an anthropological approach to study the culture of the scientist working in a laboratory (Latour & Woolgar, 1986) or to study the mediating role of medical records in hospitals (Isah & Byström, 2020). The concept of *case* includes a diversity of understandings and emphases of approach among different theorists. The use of overlapping understandings of what a case study entails has been shown to create confusion about what it really is or should be: for instance it can be the process of *conducting a case study*, the case can be a *a unit of analysis*, or it can be the *product* of a case study investigation (Merriam & Tisdell, 2015, p. 37).

For Yin, case study is an empirical method that "investigates a contemporary phenomenon (the "case") in depth and within its real-world context" (Yin, 2018, p. 15). This perspective underscores the scope of the methodology, which is to arrive at an understanding through an in-depth study of something. This understanding is reached by incorporating context, which is explicitly assumed to be relevant to interpretation. Because it is not easy to separate context from phenomenon, Yin lists three additional interrelated methodological features as being relevant characteristics of case study. These characteristics indicate that the case study method has its own logic that shapes most aspects of the study. Because the case study involves "more variables than data points", it uses "prior development of theoretical propositions to guide design, data collection, and analysis" and "relies on multiple sources of evidence" (Yin, 2018, p. 15). The latter characteristic is critical in case studies for providing an exhaustive description of the origins and evolution of the case over time (Padgett, 2017). Multiple data sources are also needed to gain an understanding of the larger context of social, political,

economic, and cultural conditions surrounding the case. The sources are then integrated by sharing research questions and, in the best cases, they complement each other.

In the case study, the focus of interest is also reflected in the research questions on which the study's research design is based. Case studies are considered appropriate for explaining social phenomena by asking questions focusing on *how* and *why*. This form of inquiry aims to trace processes of how and why something occurs rather than describing frequencies (Yin, 2018). When used in an explanatory manner, case study can also be applied when asking *what*-questions. Another way of putting it is that a case study is a form of qualitative research design that is useful when the scope of a study is focused on depth and learning (Flyvbjerg, 2006).

Some researchers focuses more on the case as a unit of analysis; for Stake, for instance, the case is the one among others on which we concentrate, searching for "the particularity and complexity" (Stake, 1995, p. xi). The author describes the case itself as an entity, "an integrated system" with "a boundary and working parts" (Stake, 1995, p. 2). It is characterised by specificity and boundedness; without boundedness, a phenomenon cannot be a case. While Stake is less specific than Yin about what separates the system from the context, these ideas nevertheless do emphasise the case's "embeddedness and interaction with its contexts" (Stake, 1995, p. 16); the meaning of contexts must be considered when interpreting a case.

At present, case study is considered a valuable methodology in itself. In the past, it was viewed more of a method to initiate the research process (for example, to generate hypotheses) if it was considered a formal research method at all. However, the use of case studies is still debated, especially in relation to issues of representation and generalisation. Questions arise about how findings from just one case can be generalised and considered representative of a larger number of cases. This form of criticism has been largely derived from comparison with quantitative methods and contested by arguing that generalisation is not the only evidence of, or contribution to, scientific success (Flyvbjerg, 2006; Silverman, 2017). To argue for generalisation, in Stake's view, would be to misunderstand the goal: "We do not study a case primarily to understand other cases. Our first obligation is to understand this one case" (Stake, 1995, p. 4). Stake adds that understanding the one case, as evidenced by repeated activities or counterexamples, for example, can help refine or modify our previous knowledge.

Yin's and Stake's descriptions of case study methodology each contain aspects that can describe the approach taken in the present study. Their ideas correspond to the motivation and scope that guide this research, which is to learn about and understand how data and data

sharing are enabled in the context of data-sharing initiatives. Data are collected from a variety of sources in order to observe this contemporary phenomenon in a particular real-world system, formed here by a research group. The system is bounded by a formal contract as well as by the active engagement of the researchers involved in the joint enterprise of collaborative research. In this way, a rich and multidimensional picture can be drawn of the particulars of the case, its many variables and working parts. The larger context is considered by including the social, political, economic, and cultural conditions that surround the research group and therefore may be significant for how the researchers act. Since not all of the many parameters and details offered by the case and its relationships to the context can be examined, the most important parts of the material to be interpreted must be selected. The results will contribute to insights and learning experiences that may be transferable to other contexts even if generalisation is not the main purpose. I agree with Latour's description of his research intentions in the well-known case study *Laboratory Life*: "Without claiming to have given an exhaustive description of the activities of all like-minded practitioners, we aim to provide a monograph of ethnographic investigation of one specific group of scientists /.../ of scientific activity obtained in *a particular setting* [emphasis in original]" (Latour & Woolgar, 1986, p. 28).

In reference to the quote from Latour and Woolgar, I conclude that this study is in part similar to and inspired by an ethnographic study. The intent was to study a research group and its social world and to be in the field as much as practically possible: to spend a lot of time with the researchers, to make direct observations of what is said and done, and to capture patterns, tacit rules, and taken-for-granted understandings. The ethnographic stance of taking a fully *holistic* perspective was obviously limited by the conceptual framework. An even more important inspiration was the non-prescriptive stance of the methodology (Padgett, 2017), interpreted as the application of an open-ended, flexible, and discovery-oriented approach.

To select a case

In concentrating on *one* bounded entity, how this case is selected is clearly of critical importance, and several issues should be kept in mind during the selection process. Primarily, the case chosen should enable maximised learning and best lead to understandings or modified generalisations (Stake, 1995). The case should be accessible and include people willing to answer questions. Stake also relates the selection process to the form of interest for

the case, as the particular case will have implications for the methods that can, should, and will be used. An investigator driven by intrinsic interest in a case and its particularity is curious to find out more about it. On the other hand, an instrumental interest is driven instead by a secondary purpose, for example a will to learn more about a research question, and belief learning may be achieved by studying the specific case. The interest in the case itself is subordinate to the research question (Stake, 1995).

Here, this single case study is mainly guided by an instrumental interest. The specific case was chosen because it was assumed to be a context in which certain processes occur and would therefore illustrate the research questions. The social environment met two main criteria of interest. First, it was an interdisciplinary community that was working with data sharing. Second, it was bound to follow the requirements of a data policy from an external party. Additionally, timing was important for observing the community's response to the funder's demands for data sharing; the group had to be currently involved in responding to the requirements and take varying kinds of actions regarding data sharing in order to deal with the externally-imposed demands.

The criterion that the research group should be interdisciplinary was based on previous research results that used various disciplines as an analytical framework. These studies showed that practices related to data differ among disciplines (e.g. Tenopir et al., 2015; Womack, 2015), which was the reason for wanting to explore more closely how data were being shared within a community formed by collaborating researchers of different disciplines, assuming the data sharing across disciplinary boundaries would create tension.[5] The second criterion was based on the fact that research funding is used as an instrument for enacting governance and changing behaviours (Jacob & Hellström, 2018), which evidently is the foundational intention of institutions adopting data policies. I wished to investigate how a research community negotiated these requirements and formed a response in order to understand what happens when policy and community meet.

In the process of looking for a suitable research group, a research librarian put me in contact with a research project that fulfilled the criteria.[6] Because of their Horizon 2020 funding, which included the Open Research Data Pilot, the group would have to comply with

---

[5] Quite early in the research process, when I was spending time with the project group members, it appeared that there were very few tensions in the community. Because collaboration in the project ran rather smoothly, as did the sharing of data between participants of different disciplines, the material gave no further analytical ground to work with, and therefore part of the original problem formulation had to be abandoned. This result will be further discussed in chapter five.

[6] This librarian had been assisting the project group when they developed their data management plan. Therefore, the librarian was later interviewed for this study.

demands concerning data. These demands will be presented in detail after this section together with detailed background information about the chosen research group. An initial meeting with the group's project manager in June 2017 resulted in a formal request for collaboration to the management group. The application included explanations of the project's aim and the intention to conduct interviews with the project members. Shortly thereafter, the project management invited me to attend a project workshop where I presented the study's focus at a poster session and the request for collaboration was eventually granted.

Gaining entry to the field and finding people willing to participate is often critical for a research project and the project manager (interviewed as a project member) became a person crucial for this investigation's data collection. Since our first meeting, we remained in contact throughout the thesis work. This person became a stable link between me and the group and was in the true sense of the word a *gatekeeper* (Padgett, 2017, p. 71) of the research community. The project manager ensured that the interviews and observations came about; the researchers in the group were not easy to get hold of. Besides agreeing to a formal interview, the project manager (and to some extent also the co-coordinator) on various occasions took the time to explain things about the context that, to me, was a completely new environment and research area, and they answered questions on all matters. They handed me important documents, emails, and project presentations that all became important evidence to use in this study.

The role of the gatekeeper is a power position that can be used both to provide or limit the material that a researcher gets access to (O'Reilly, 2009). I was dependent on this person through whom I got access to much information. At the same time, it was up to this researcher to decide which information were to be shared. It is difficult to judge whether this person denied me any important material. However, because the project manager (and most other project participants when I contacted them) always accommodated my request for information, in my judgement, her/his role as gatekeeper did not hinder me from accessing necessary information.

Because research projects often last for several years and doctoral project time is limited, it was not possible to follow the entire process from start to finish. Therefore, the study focused on the initial phases of research project work; that is, data collection was mainly carried out during the first and second year of the project's five-and-a-half year period. During this time, the group's first encounters with the data policy took place and the data management plan was developed and delivered. Project research was up and running, data were being shared intensively within the group, results were published and some of the generated data were

made openly accessible in a variety of ways. This project period was however not a time when research work was concluded or when all community data was compiled, revised, and prepared for final storage, or when work processes were finally evaluated. For this reason, the investigation did not follow up on to what extent the group followed their initial intentions to make the data openly accessible, as formulated in the data management plan, at the end of the project.

The European Commission's requirements about research data

Research projects included in the Open Research Data Pilot must report how they will manage research data generated and/or collected during the project when the initial application for funding is made. In the application, issues must be clarified regarding what data types will be generated, what the standards are and how they will be used, how data will be exploited or made accessible and, if not, explanations for why not. However, the pilot is flexible, and projects have the possibility to opt out by providing a justification for doing so, for instance around privacy concerns or Intellectual Property Rights.

The pilot contains two pillars: developing a Data Management Plan (DMP) and, if possible, providing open access to research data. Funded projects must adhere to the following conditions:

- Develop (and keep up-to-date) a Data Management Plan.
- Deposit your data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate your data
- Provide related information and identify (or provide) the tools needed to use the raw data to validate your research (OpenAIRE, 2017).

Open access to research data is defined as "the right to access and reuse digital research data under the terms and conditions set out in the Grant Agreement" (European Commission, n.d.-a). The EC writes that data management plans "are a *key* element of good data management" [italics in the original], describing "the data management life cycle for the data to be collected, processed and/or generated" (European Commission, n.d.-a). The funder recommends and provides a template for the DMP, use of which is voluntary. The Open Research Data Pilot is intended to apply primarily for "the data needed to validate the results

presented in scientific publications." To make other data openly accessible is voluntary. The European Commission defines research data as "information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion or calculation" (European Commission, n.d.-a).

Granted projects must deliver the data management plan within the first six months of the project period. However, the data management plan should be dynamic and be updated during the project life cycle. During the funding period, projects may opt out of the pilot at any stage. The Commission's stated goal is that research projects follow the principle "as open as possible, as closed as necessary" when making research data accessible and difficulties as security, privacy concerns or Intellectual Property Rights are taken into account (European Commission, n.d.-a). Granted projects can claim costs related to making data openly accessible. The funder does not state whether the DMP will be submitted to review in any form. It is however explained that the plan needs to be updated in time for the final review at the latest if this was not carried out during the project's duration.

The case – a growing interdisciplinary community within a new STEM area

The studied research group consisted by 18 persons at project start in January 2017. At the first conference held by the project, the project co-ordinator described the group as a consortium or a community, and the same terms are used on the project's website. The research group is formed by six individual partner groups, or *beneficiaries* according to the European Commission vocabulary. These comprise four universities, a private research institute, and a small research-intensive enterprise. The partners are spread out over four countries; three are member states of the European Union while the fourth is an EU associated state. The international group consists of researchers of varying seniority levels, doctoral students, and postdoctoral researchers. The research project is currently funded by the EU and has a budget of over 6 million € over a five and a half-year period. Funding is received gradually after *work packages* have been delivered, and the last part of the funding is received after the final report is approved. This assessment is carried out by external experts that review whether the project has achieved the established aims.

The topic of the research project lies within the Science, Technology, Engineering and Mathematics (STEM) disciplines and is relatively new; the community that has formed around it has existed for just over ten years. The group work with developing new technology by using a set of methods in a unique manner. According to the researchers, the central idea of

the research emerged from experimental results found around ten years ago. At this time, some of the current project members had been working together for a period and decided to submit the first research proposal in 2006. To this core group, new members have been added since then. As a result, which was commented by the researchers, some of the project members know each other very well.

On the project's website and by a few of the researchers themselves, the community is described as collaborative and highly interdisciplinary; four disciplines are represented in the group. Its current composition is a result of an ongoing search for persons with suitable skills and specialities to succeed in solving the present research problems. The project management consider the representation of multiple competence as crucial for being able to move the research forward. To attract a larger scientific community is one of the two main goals, beside developing the technology, their website state. For making the community grow organically, the project's strategy is to continuously search for researchers with relevant qualifications that are found lacking in the group as work progresses. Several researchers said they had joined the project after having been offered participation at conferences and international meetings. For all researchers, this project is one of many where they are engaged in research work, in parallel with other projects or in their home institutions. At workshops, the group discussed solutions for attracting researchers with needed competence and the innovation award the website announce is one way to establish contact with interested scientists or students who could contribute with novel ideas.

A project management team formed by a co-ordinator, a co-coordinator and a project manager, representatives from two partner groups, lead the project (see Figure 1). The project organisation also has a larger management group called the General Assembly where representatives from all project partners are included. The project's six partners each include one to five researchers. Most of them are composed by researchers from different disciplines. Research tasks are organised into work packages and *deliverables* (distinct project outputs) This structure is in line with the funder's requirements on how research must be structured and presented in the application template. Depending on their expertise, the researchers are involved in different work packages. This means that a researcher can be engaged in work within several work packages and thereby be involved in different constellations within the group. The work organisation was frequently mentioned in both interviews and at observations. For example, researchers spoke of how communication was carried out on different levels such as local, project or work package level.
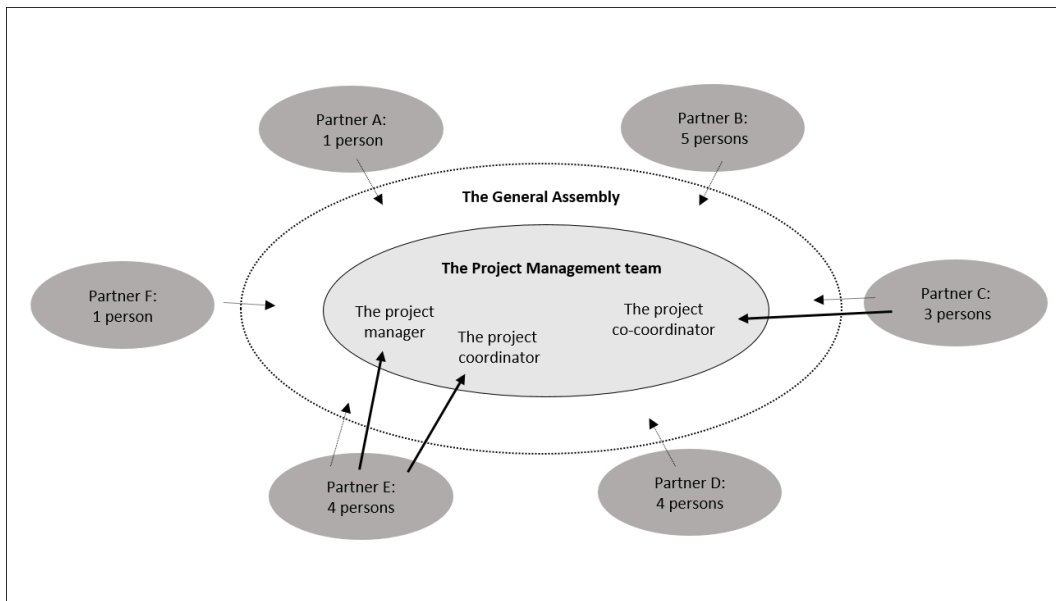
Figure 1 represents the six partner groups and their number of members. All partner groups have representatives in the General Assembly and two groups are represented in the Project Management team.

To organise communication between members distributed geographically, the project members actively work according to a detailed communication plan. At times, researchers referred to it at project meetings and interviews. This plan structures the frequencies and forms for meetings and keep the group together. For instance, it established that researchers working on the same work package should have short informal meetings virtually every second week, and that the whole consortium met in person at project meetings scheduled every ninth month. As a common infrastructure or tool for communicating and sharing data and information, the project mainly used an implemented online cloud service to which all members had access and could edit. (This will in the forthcoming text be referred to as *the file sharing solution*).

It is difficult to describe the degree to which collaboration takes place between the partners in the project. Some partners appeared to have more frequent contact than others, but this varied over time. A presentation slide titled "Workflow and data generation" written by the project show how the group chose to illustrate the project partners and the workflow between them (see Figure 2). It also displays how data are generated and passed between the partner groups; occasionally in a linear way, as between partner A, B, and F in the beginning of the flow chart, occasionally back and forth in an iterative way as the double headed arrows between for instance C and E or B and E indicate. Group B and C appear more than one time and are involved in varying phases of the research process. This illustration corresponds quite

well with the impression I got of how the partner groups collaborated at observations. Partner A and F were the more theoretical researchers. They originated the processes and were less directly involved later in the more experimental processes that followed even if they were consulted along the way and were active when discussing how to move forward with the experimental results. Since not all partner groups did not collaborate directly, all of them are indispensable to develop the research.

To sum up, the case consists of a heterogenic collection of individuals in that it is formed by researchers representing different disciplines, nationalities, workplaces, and seniority levels. The group has collaborated in similar formal constellations and has been granted shared funding before. It is dynamic and growing which entail that a few members are recently added to the group as newcomers while others started collaborating on this subject ten to fifteen years ago. Though the group is of a distributed nature, with researchers scattered over four European countries, the researchers are frequently communicating with each other and publish scientific articles together. Over the years, the members have formed relationships and shared ways of engaging in their common research activities including organised information flows and shared resources.
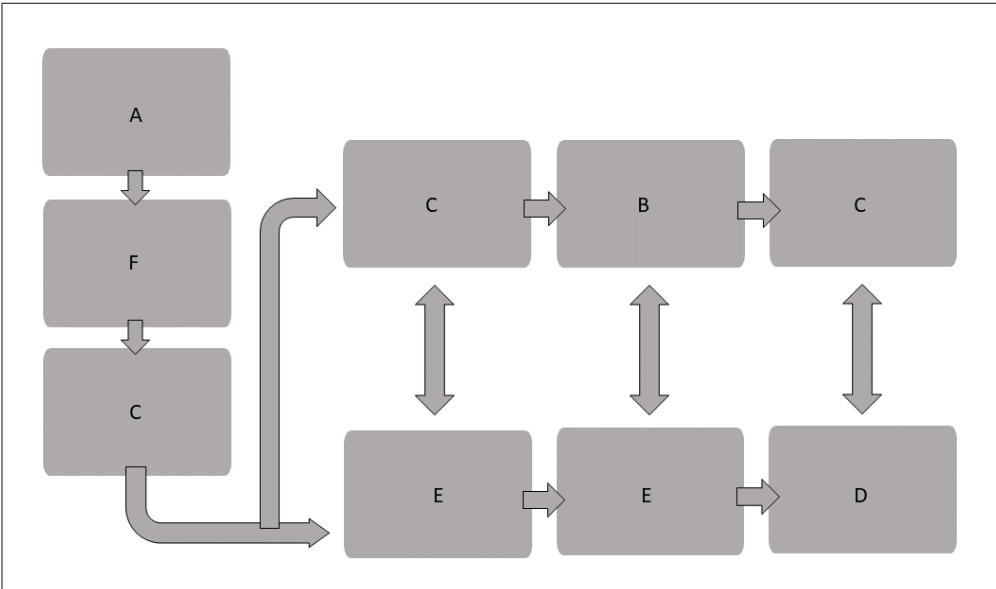
Figure 2 represents how the group illustrate workflow between partner groups and data generation within project on a presentation slide.

Methods of data collection

The material for this investigation consists of interviews, observations in various settings, and documents. Additionally, photographs, email conversations and other documents, publications and a small survey have been used as completing material. How these are used and how they integrate is presented here together with ethical considerations that emerged and the development of analytical process.

Observations and note taking

The case, the research group in focus for this study, consisted of 18 persons. Because the researchers were scattered over four European countries, observations had to be carried out in accordance with the group's meeting plan. Already at the start of the project, the group had scheduled various kinds of gatherings for the coming five-year project period. All project members would meet in person once or twice a year, and these meetings were the occasions on which I had the opportunity to do observations of the group as a whole. The aim for all observations was to gain maximum insight of the ongoing activities in the group. The observations, or the *in situ* monitoring, were carried out on three different occasions at gatherings of different types in different countries, over the period of more than a year. At all observations, I took notes on paper, jotting down what was relevant for the observation focus of that time. More informal chats with the researchers also became important material. These talks were noted but not recorded.

The first observation took place on a two-day start-up workshop in September 2017 (9 months into project) where both project researchers and researchers either loosely connected to the project or working with adjacent subjects participated. The scope of the workshop was to get people to meet and exchange information about projects they were working on. These first personal encounters with the researchers, listening and observing varying presentations on the topic, introduced me to the research area broadly and to the specific project, the problem the group would try to solve, the data types they generated, and their group organisation. The focus for my observations was to understand what data were produced by the different disciplines, and how data was shared among the group partners, which sometimes included sharing across disciplinary boundaries. The poster I presented at this workshop served as an introduction of myself and my investigation and facilitated contact.

After this meeting, the project manager sent me a list of the project's participants so that I could start contacting them to schedule interviews.

For the second observation, I attended a one-day meeting in March 2018 (15 months into project) held only for project members. The meeting had the character of a work meeting and was arranged for preparing the presentations of the first project year's research results to representatives of the funding agency, and 15 of the 18 researchers participated. On this occasion I sat among the researchers and observed the group working and collaborating directly. I learned more about the topic and methods, and of the relationships between the researchers, what they agreed about, their roles, their tensions. The tasks of each researcher also became clearer, as did which researchers and partner groups communicated and shared data more frequently, and how this was carried out. This meeting was also an occasion for observing whether, and if so how, the data policy played a role in their work when meeting. The knowledge I had gained by starting the interviews three months earlier enabled me to better understand the observations I made, to see repeated patterns and features in the researchers' activities. At the same time, the observations contextualised the content of the interviews; I could fit individual or group statements into the relations of the larger group. I thus learned more for the coming interviews and for analysing them.

At this point I was no longer a stranger to the researchers. At the project website, I was officially described as an associated project partner. I saw this as a sign of inclusion and possibly a method for increasing the interdisciplinary aspect of the project, which was something I was explicitly striving for. Attending this meeting, chatting at coffee breaks, and trying to blend in as a more familiar person helped create a common experience which the researchers would later refer to in interviews or when meeting. Informal talks with the researchers also provided information that complemented the information from the more formal interviews. I had conducted interviews with several of the researchers and they had gotten to know me and my research interests more. It was my impression that the participants had become accustomed to my presence and that my observations did not cause any changes in the activities they were engaged in.

A third observation was carried out in November 2018 (nearly two years after the project start) when the research group gathered for a two-day meeting. The purpose of this gathering was to present what had been achieved within each work package since the last physical meeting and plan future work. My aim for this observation was to gain more details of data and data activities: how the researchers talked of data and sharing, how they presented and described data to each other, and what were their common understandings. On this occasion,

new researchers had been added as members and there were 27 attending participants. These new members were not interviewed, but they were included in the observations and talked to informally.

Eventually, I also had the opportunity to do observations in two laboratories in two different countries. The researchers connected to these laboratories guided me in the environment where they constructed apparatuses and carried out experiments. They showed the many forms of equipment and their uses, and performed an actual experiment in front of me. The laboratory visits were unique occasions to directly observe the researchers on their home grounds, handling the machines and generating experimental data. I documented via jotting down notes and taking pictures of instruments and data.

Interviews and transcriptions

The first interview took place in November 2017, not long after the first observation, and interviews continued until July 2018. This means that interviews were mainly carried out during the second year of the five and a half-year duration of the research project I studied (2017–2022). This period was a highly active project phase. The project had passed from a stage of performing experiments locally into more active collaboration which included frequent sharing of data among partners and publishing results based on project research in articles. When starting the interviews, I had met the researchers once at their start-up workshop. We were thus acquainted, and they had been informed of my research interest.

Altogether eight interviews were conducted with 16 of the 18 original project members that were included at project start in 2017. The researchers were asked by email to participate. Participation was voluntary and two persons declined. In addition to the researcher interviews, an interview was carried out with the librarian who had assisted the group in the initial phase when they developed the project's data management plan. This person's account of the interaction with the project manager provided details of the concrete issues the group needed external help with during the process of forming a data management plan and making data openly accessible.

The interviews were conducted either in face-to-face meetings or online. My intention had been to arrange physical meetings, when possible, but that proved to be difficult to organise. The geographical dispersal of the informants meant that nearby researchers were interviewed in their work settings or at a library, while researchers further away were interviewed via video link; this was also according to what best suited the participants. With one exception,

the researchers of the four partner groups were contacted directly. The group leader of the fifth partner group wished to be contacted directly in order to be able to organise and schedule the group colleagues. This organisation thereby came to decide that these researchers would be interviewed together. In total, five of the researcher interviews were made individually and three in groups or pairs. The interviews lasted between 40 and 90 minutes. One larger group was interviewed twice since the interactive conversation in this group made the time for the first interview insufficient. All interviews were audio recorded after oral approval by the interviewees had been obtained.

The interview material for this study is thus composed of a variety of individual interviews and group interviews. There are individual stories of the data work and statements expressed in the setting of a group. Dialogue between researchers frequently occurred during the group interviews, which was an advantage of having group conversations. A group may provide prompts to talk because the informants can comment and respond to each other and be stimulated by having an audience other than the interviewer (Macnaghten & Myers, 2007). Group interviewing can also stimulate dynamic processes that encourages argumentation and questioning, which subsequently can trigger otherwise implicit knowledge (Justesen & Mik-Meyer, 2012). Having group interviews thus helped shape the outcome since the researchers commented or completed their colleagues' statements with additional information. Group interviews resulted in more detailed accounts on matters that particularly concerned the researchers as the group could influence more the direction of the discussion. The effect can be clearly seen in the fact that some issues in the presentation of the results are overrepresented by one partner group.

Drawing on previous research of the subject, the knowledge I had gained of the research group up to this point, and the research questions, an interview guide was developed (Appendix I). Experiences of how the questions were responded to or whether they stimulated interesting discussions or not guided the succeeding interviews, and the interview guide was slightly modified. As an introduction, the researchers were asked a few questions about their discipline and how they had become involved in the project. Then, all interviews started from an example of data I had asked the researchers to bring to the interview beforehand. The researchers showed illustrations of data that they had published in scientific articles, or displayed still or moving images from electron microscopes. At an interview in person, the group showed a transparent paper floating in liquid in a small laboratory dish. One researcher opened a document containing graphical representations of networks s/he had built for

developing software. All the examples of data will be presented further in detail in chapter four.

This idea to ask researchers bring data was based on several assumptions and hopes. Activities could be difficult to describe in words, as these doings might be an obvious part of the daily work. Talking to the researchers during the first observation, I learned that asking questions about unspecified data, as detached from context or as isolated things, was not fruitful – it only resulted in questions and confusion. In addition, the researchers' choice of data to present was informative in itself, and the request to bring data appeared to have started reflections around data on behalf of the researcher even before the interview. Finally, it was hoped that having researchers speak about data that they themselves had produced could evoke a feeling of being on home ground in what might otherwise be experienced as an uncomfortable interview situation. I consider the strategy was successful and worth re-using for other studies. Conversation started easily, which was favourable for the ongoing interview and made the researchers move on to tell of other data-related issues. In some cases, however, the question resulted in lengthy explanations about instruments and methods that were difficult to follow without expertise in the field.

After discussing the researchers' data samples, a number of semi-structured questions on two themes were posed. The first theme regarded how data were produced and communicated with project partners. The second theme aimed at the process of encountering the data policy, the development of the data management plan, and the role data policy played for everyday data practices. The interview guide was used as a support more than rigorously followed, which meant that the interviews took slightly different directions. It was important to let the researchers' responses decide what questions were to be asked (Holstein & Gubrium, 1995). I wanted to make space for the informants to elaborate on issues that engaged *them* and for me to follow up interesting or unexpected traces. Much later during analysis, it became clear that certain issues had not been followed up in the interviews, or further questions were raised, and I then contacted the co-coordinator. S/he answered these questions via two email conversations in June 2021.

All interviews except one were carried out in English. The one interview in Swedish, I translated myself. Most researchers in the project did not have English as native language, nor do I. Thus, most of us had to express ourselves in a language secondary to us, a language use which may have led to some loss of nuance or sliding of significance. The interviews were transcribed verbatim, as in case studies it is particularly important to understand everything about the research subjects' experiences (Pickard, 2017). The level of detail of transcriptions

is a much debated issue since early decisions about transcribing impact the analysis (Tracy, 2019). As the researchers' experiences are the focus of this thesis, I aimed to reproduce the registered conversations as accurately as possible, including pauses, laughter, hesitations, and emphases. Evidently, too many grammatical errors and discourse markers (so, like, also) may hinder the reader's comprehension. Quotes used in the thesis have therefore been gently edited or shortened when necessary. In quotes, a researcher's emphasis is presented in italics.

Interview 1: one researcher, library, 30 November 2017, 53 minutes

Interview 2: four researchers, their institution, 19 December 2017, 40 minutes

Interview 3: five researchers, Skype, 19 January 2018, 54 minutes

Interview 4: one researcher, her/his institution, 29 January 2018, 90 minutes

Interview 5: follow-up for interview 3, three researchers, Skype, 2 February 2018, 40 minutes

Interview 6: two researchers, Skype, 12 April 2018, 75 minutes

Interview 7: one researcher, Skype, 10 May 2018, 47 minutes

Interview 8: one researcher, Skype, 28 June 2018, 80 minutes

Interview 9: one librarian, her/his institution, 6 July 2018, 45 minutes

Documents and archival materials

Different types of documents and archival data has also been used as empirical material for this investigation study. A particularly important document is the project group's data management plan, titled "[Project name], Data Management Plan".[7] The data management plan is a detailed description of the data types of the project, strategies for handling data and making them accessible as well as the resources and cost it entails. The project group I studied drew the outlines for this plan as part of the research proposal. I have taken part of these as well as an early version of the developing data management plan that included the project researchers' comments. As these also have served as material, it is however mainly the final version of the data management plan that was delivered to the funder six months after project start that was analysed thoroughly here.

The information in documents regards certain circumstances which are noted at a specific time and place. They tell about the authors' perspectives and how they want to present reality. Contrary to interviews, documents are produced before researchers' interaction with subject why they therefore exist independently of the investigating researcher (Johannessen et al.,

---

[7] As a project outcome or deliverable, the data management plan document is openly available at the Community Research and Development Information Service (CORDIS) website for EU research results.

2020). Their lack of reactivity is an advantage as they therefore can be seen as representing voices directly and unfiltered (Padgett, 2017). However, this naturalness should be considered, and its relative quality, accuracy and completeness has to be evaluated. The investigator must keep in mind that authors may have various reasons for omitting certain issues when they produce documents.4

The project's data management plan is seen as a document that describes how the project members wanted to present their future data management activities during the project period. Formally, the data management plan is a deliverable. It should be sent to the funder who is one intended recipient. The researchers themselves are also intended recipients, since the funders' intention is that the data management plan should also serve them; the plan is to be a "living document" (European Commission, n.d.-a) that helps researchers make data FAIR[8] and well managed. The fact that the project's funder is a receiver has however to be considered when using the data management plan as material. It constitutes a part of the group's shared repertoire and make agreements, interpretations, and expectations, regarding their structured data management visible and explicit. But the funder is in a power position in relation to the project, and thereby, being a receiver of the data management plan, the document must be read as potentially aimed to comply with the funders' requirements. Thus, the plan may be a statement of how the researchers intend to handle the project data during the five-and-a-half-year period, and at the same time also be a statement that reflects what the group supposes the funder wants to see. Putting the data management plan in relation to the interviews and the other material, makes it possible to compare varying descriptions of activities. This has thus been one method of a reaching a more nuanced and accurate perspective on the data management plan's descriptions.

The data management plan consists of a seven pages long document divided into six headlines as recommended by the Online Manual (European Commission, n.d.-a). The first two, the Data summary and the FAIR data, together take up six of the document's pages.

1. Data summary

2. FAIR data

3. Allocation of resources

4. Data security

5. Ethical aspects

6. Other

---

[8] FAIR is an acronym for Findable, Accessible, Interoperable and Re-usable, a widely-used standard for how data should be made openly accessible.

I got access to the data management plan by the project manager in January 2018 when I had just started interviewing the researchers. The final plan was presented to the funder along with other project results at a review in March 2018.

I have also taken part of documents related to the development of the data management plan. Some of them were referred to in the interviews by the researchers. Examples were parts regarding research data from the projects' research application, presentation slides of the project's data structure and organisation structure

Other archival data used were the project's webpage, the email correspondence between the project manager and the supporting librarian, and the research community's publications in journals (observed in February 2019), and openly shared data (following links in journal publications). The journal articles were used for investigating if these contained accessible research data as supplementary information. In the cases when these data were source code data, articles linked them to GitHub[9] where they were observed. Similarly, the digital data repository Zenodo[10] was used to observe how the community had made used of it during project.

Images and photographs

Different kinds of images have contributed as and memory support to the interviews, field notes and documents. These are photographs I took of the data samples that researchers presented to me at the interviews as well as photos of the group taken at observations showing the researchers, presentation slides, instruments, machines and experiments. A couple of researchers also emailed me around ten images of data they worked with.

Survey

The collected material also includes a small survey (Appendix II). It was sent to the researchers by email in November 2018, just before the occasion of the last observation when the whole research group would gather. At that time the group was almost two years into the project period. The research group had published several articles that were displayed on the project's website. I assumed the researchers had made data openly accessible as well. The

---

[9] GitHub is an online platform that offers services for software development processes. It can be used for building and storing of source code.

[10] Zenodo is a general open-access repository, described as "a catch-all repository for EC funded research" developed under the European OpenAIRE programme (CERN Data Centre & Invenio, n.d.). In the repository, researchers can archive journal articles, datasets and other research related digital material.

purpose of the survey was thus to get a status report of whether any of the project data had been made openly accessible until this point. The intention was to enrich the picture and understanding of the perceptions towards data sharing and activities by relating the survey to the other collected material.

The survey posed one question, "Have you until now made [project name] data openly accessible, formally or informally?" and was answered by yes or no boxes. It was possible to also specify *where* the data had been made accessible. At first, the survey resulted in few responses, partly due to institutional firewalls. However, at the meeting, all researchers were given the opportunity to fill in a printed version of the form. In the end, 15 researchers of the original list of participants responded the survey, including one of the persons that had declined an interview. One of the previously interviewed researchers had left the project and the project manager declined the survey since s/he did not generate any data.

Ethical considerations

Because audio recordings and photographs of persons are considered personal data, this study was reported to the Norwegian Centre for Research Data (NSD) in November 2017 (project number 56717). The data are regulated by the Norwegian Personal Data Act and are not considered sensitive. The participants of the research group were informed of the case study's interest and activities, as collecting data would entail an invasion of the group's space (Stake, 1995). The group was first informed via the General Assembly (the project's larger management group with representatives from all project partners) and second via email when each member was personally asked to participate in the study. The information clarified that participation in the study was voluntary and that anonymity principles would be applied; the collected data would not be shared with anyone other than my supervisors and the results would be presented in such a manner that the project and the participants could not be identified. Later, when the interviews were conducted, the informants all gave oral consent to audio recording.

The research group chosen for this case was not an altogether easily accessed community. Geographical distances and researchers with little time complicated data collection and limited the occasions on which direct observations were possible. Besides the formal approval from the General Assembly that was needed to enable collaboration with the group, several additional confidentiality agreements had to be signed during the study by the six

international partners in order to grant me permission to participate in the group's meetings. I was permitted to participate in all meetings that took place during my observation period except for two review meetings between the funder and the research group. Before consenting to be interviewed for my project, one partner group asked for additional specifications about my study's scientific purpose and method (in addition to the information outlined above). This was the group that worked at a private research institute who were used to working under highly restricted contract conditions with industrial partners.

Regarding the promised anonymity, the decision about whether to anonymise the project, the researchers and their disciplines was not straightforward. It was necessary to find a balanced way to present the results in a fruitful and representative form that would not violate the promise of anonymity. Because the project co-ordinator gave her/his consent for declaring the project's name, the choice to reveal it up was up to me and the decision would have implications for the investigation. If the research project's name was revealed, the researchers could be identified. Therefor using the name was out of the question, even though the issue could have been solved by for instance asking each researcher for consent later in the process.

To name the project and more importantly all the disciplines would have allowed specific details regarding disciplinary differences in data practices to be related to previous knowledge of the different practices of the different disciplines. Doing so would have had the advantage of making it easier for readers to follow the reasoning. To present the results without naming disciplines of the researchers might make the results appear indeterminate and vague. In the end I chose not to reveal the project name and only mention that researchers were engaged within the Science, Technology, Engineering and Mathematics disciplines. I did not want to take the risk of full identification and exposure of the informants in the group. To name the project would complicate the intention to present rich and detailed descriptions of the material. To produce results translatable to other contexts was in the end not within the scope of this case study; instead, the focus in on understanding this specific case and this community's ways of understanding data and sharing data.

Furthermore, as mentioned previously, my first encounters with the group also shifted my focus of interest, because the material did not provide the expected analytical ground to continue with the initial questions. Disciplinary differences turned out to be less important; practices related to other identities, such as experimentalists and theoreticians, turned out to be a more significant issue to the researchers. The importance of presenting the specific disciplines was therefore lessened, with however two exceptions. Since the data of the two theoreticians revealed interesting aspects related to their specific disciplines' methods and

their research approaches, and since identification would not be possible, I decided to reveal their disciplines.[11]

As a result of these choices about presenting the results, several precautions were taken when describing the collected data. The transcriptions were anonymized to prevent identification of individuals and institutions directly or via triangulation. Personal information such as workplace, geographical location, and disciplines (with the exception of the mathematician and the computer scientist) were erased. Quotes taken from the project's data management plan and websites were first Googled to ensure they did not enable identification of the project. The photographs I took at meetings and in laboratories were taken and used with permission. The majority of the photographs of the project data included in this thesis were provided by the project co-coordinator with her/his consent to present them. As for the project's methods, technology, instruments, and data file formats, they are to some extent described in a way that could potentially enable persons with certain skills to discern what form of project or institutions are involved. Even if the responsibility for the collected data was mine, as a doctoral student, NSD brought attention to the fact that for this specific study, part of the responsibility also lay on the informants. As researchers, they were responsible for the information they shared with me regarding their research. On their behalf, having agreed as committed project participants to professional secrecy concerning their ongoing research, they were responsible for what information they chose to reveal to me.

Throughout the thesis and in quoted material from interviews, I combine genders to ensure that both are included equally. This means that "s/he" is used as singular noun, and in addition "her/him", "hers/his", and "her/himself". A fully ungendered pronoun option would have been preferable since gender is not of relevance for this investigation. However, using "they" as a singular pronoun would have obscured the number of informants that are referred to, something that has to be clear in this thesis, why the other model of combining gender for pronouns was preferred.

When presenting the results, each informant was given a pseudonym (A-F) representing the six partner groups. The librarian had a separate pseudonym (G). For the project manager (E1) and the project co-coordinator (C1), pseudonym and role title are used interchangeably for stylistic reasons. The same method is used for the mathematician (A) and computer scientist (F). E3 is the pseudonym of the project co-ordinator; this person did not participate in

---

[11] This decision will be revisited in chapter five, the final critical reflections of this thesis.

interviews, but is quoted when speaking in a meeting. Two additional researchers (Researcher 1 and Researcher 2) did not participate in interviews because they joined the project at a later stage, but were engaged in dialogue at a workshop. It is not clear which partner group they belonged to.

A: the mathematician (from the enterprise)

B: private research institute (1-5)

C: university (1-3)

D: university (1-4)

E: university

F: the computer scientist (from a university)

G: the librarian

Researcher 1

Researcher 2

Analysis

The question of when exactly one begins to analyse, instead of merely collect, empirical material is debatable, but on the whole I agree with Stake, who claims that analysis begins as soon as the researcher enters the field. When meeting the research group, the case, for the first time, I instantly started trying to make sense of it, to make individuals and their activities understandable. The process of analysis is to take a "new impression apart, give meaning to the parts" (Stake, 1995, p. 71). As interpretation is central when analysing case studies, the material leaves many possibilities available to the researcher, and much responsibility to do a just selection of the case's most significant aspects. This selection process depends on the investigator's foundation in previous research, theory, and familiarity with the collected material. In this case study, certain features and particulars immediately stood out as interesting and highly relevant for answering the research questions. At the same time, many more variables in the material appeared to be important for describing context and thereby explaining the results. It was not obvious, though, how to balance these factors.

My intent was to approach the material with an open and unprejudiced attitude in order to try to understand the researchers' ways of seeing things (Fejes & Thornberg, 2015). Answers

to the research questions were to be found with the help of both the theoretical framework and related research that provides a language for describing the empirical material and communicating the results (Ahrne & Svensson, 2015). However, what constituted the relevant related research and theoretical framework was something that emerged over the course of the entire thesis work. Research questions are rarely completely formulated at the start of a project, but instead are reformulated or refined during analysis. The general analytical process is thus slowly shaped in an iterative learning process that swings back and forth between theory, collected data, research questions, and related studies: reading, interpreting, re-evaluating, describing, and explaining. As described earlier, the research questions were first reformulated shortly after first contact with the research group, as the material did not give further analytical ground to work with certain issues, and part of the original problem formulation was abandoned.[12] Similarly, after the final occasion for observation, it became obvious that what data were to the researchers was a question that needed to be addressed, which also implied a change of focus for the material analysis.

The most intensive phase of data analysis started in parallel with the last interviews. Initially, the material was read and re-studied numerous times. In the struggle to find an analytical strategy, the exploratory approach allowed me to try out methods of grounded theory, a data oriented methodology in which theory is closely tied to, or grounded in, the empirical material than in other methodologies (Alvesson & Sköldberg, 2017). The methodology's systematic yet flexible guide for analysing data to better understand social processes (Fejes & Thornberg, 2015) was appealing. Therefore, I tried to use this methodology's way of coding for structuring the material and identifying themes, and coded all transcribed interviews through substantive or open coding (Alvesson & Sköldberg, 2017). Every transcribed line was given a short summary with the aim to then form categories. This time-consuming work did not lead to the step of theoretical coding. I learned that to combine parts of different methodologies was not a fruitful way of moving forward, as different rationales collided. For instance, my study had − in line with the case study but not with grounded theory − research questions formulated beforehand that I wanted to answer. However, what I did gain from the work of coding was intense interaction with material, making me profoundly familiar with material that had initially felt difficult to overview. Another contribution from the grounded theory was the method of posing open questions

---

[12] This result will be further discussed in chapter five.

during coding: "what is happening, what are these data a study of, what do data suggest, how, when and why do the individuals' actions occur and what social actions, processes does this indicate" *et cetera* (Fejes & Thornberg, 2015). My experience was that these questions helped maintain an openness of mind and create distance from the obvious.

Concurrently, I tried other ways to orient myself within the material. As a complement to reading the interviews, I listened to them several times. Listening to the audio recordings made factors like relational aspects, emphases, and emotional expressions clearer than they were in the transcriptions, even though I had intentionally tried to include many details. In order to place focus on the researchers' activities, I extracted all verbs related to data activities, in line with the practice theory lens.

To describe the material so that the reader is able to follow the reasoning from the initial research questions to the final findings, it is essential to maintain a "chain of evidence" (Yin, 2018, p. 134), and it remained to be decided how to present the material in a way that would tell the story fairly (Pickard, 2017). As analytical strategy, I decided to put the material in chronological order (Yin, 2018) and that an illustration of the research data life cycle would be favourable for this purpose. This illustration could provide a form for structuring all the material and describe a broad yet detailed picture of how the researchers worked with data in different phases.

At one stage in the analytical process, the choice of conceptual framework was refined, and the research questions were similarly elaborated and further specified. This step narrowed the focus which made analysis progress more successfully. Narrowing the focus was necessary because the practice-based lens did not provide sufficient guidance for navigating the empirical material with its many parameters. A more articulate and practical instrument was needed for sorting and structuring. The new conceptual framework better suited the empirical material and provided a more practical analytical instrument. I used this framework as "a guide about what to pay attention to" (Wenger, 1998, p. 9) as well to identify particularly relevant issues to analyse.

Finally, a few words on the act of writing as a method in itself. Writing and formulating were central for the analysis and the structuring of the material. The emerging texts are the visible traces of the abstract and invisible research process; they are, at least to some extent, the materialised expressions of what is going on and of how work progresses. During the creative act of writing, thinking happens, which means that writing is also when research happens.

The research data life cycle

Empirical material is not readily organised. As a descriptive framework for presenting the collected data, I use an illustration of research data life cycle, which entails that the material is set up chronologically as an analytical strategy (Yin, 2018). Illustrations of research life cycles and research data life cycles are widely used by policy makers, research funders, data repositories, support services, and within Library and Information Science research to present what has become generally viewed as distinct data work phases within a research process (e.g. European Commission, n.d.-a; Tenopir et al., 2015; Tenopir et al., 2020). The research life cycle is a visualisation I have often come in contact with as a university librarian, and it was therefore a familiar form for overviewing and organising the empirical results. In relation to this investigation, I expect that using the illustration will serve for displaying both a deep and broad picture of the researchers' conceptions of data and data sharing in close relation to other related activities that include data. Furthermore, using this well-known illustration may also help make the empirical results accessible and comprehensible to a wide public familiar with the idea behind it.

The general research data life cycle is often presented as constituting a smaller part of a larger research life cycle. This research life cycle has been described as five stages, *ideas − partners − proposal writing − research process − publication* (Tenopir et al., 2011, p. 2). In this example, the data activities form their own cycle within the *research process* stage. Another widespread model of the overall research process divides the life cycles into four stages, *fund R&D − perform the research − communicate the results − apply the knowledge* (Björk, 2007, p. 13).

The specific illustration I chose to present the empirical material is thus one among many that have been developed (see Figure 3). This particular illustration has quite detailed descriptions of the elements and activities included within the various stages, which is unusual and was helpful. The illustration was used allowing for minor adjustments in the presentation (to the exact wording of the names of the stages, for example) to better accommodate this study's material. Since data was collected after the research project was initiated, the cycle's first stage (*Research idea/concept*) was left out. In the presentation that follows, I have grouped the eight different stages named in the cycle into three categories. The research group was studied mainly during its second project year, which made the empirical material particularly rich regarding the stage *Research data planning & design*, when the group wrote the data management plan, and the stage *Data collection*, where the project data

are described. The first category presented thus includes only *Research data planning*. The second category includes *Data generation*, *Data processing*, and *Data analysis*. The third category include *Data publishing & access*, *Data preservation*, and *Data reuse*. The material that concerns the final phases (*Data preservation* and *Data reuse*) were mostly from interviews that asked the question of how the researchers expected they would carry out these activities later on in the project period. As every illustration entails simplifications, and I draw methodological conclusions from having used one to present the material, this use is commented on in chapter five.
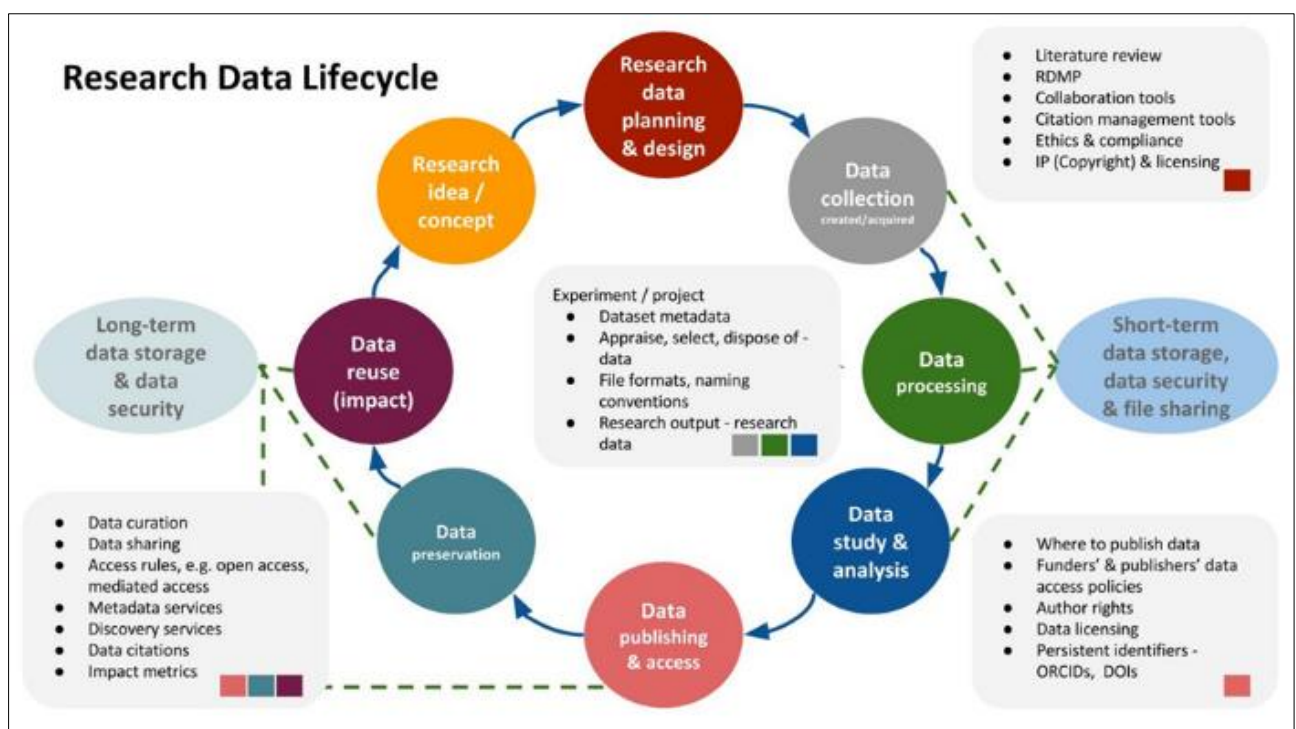
Figure 3 demonstrates the research data life cycle used for organising and presenting the empirical material of this thesis. The illustration was created by Kathryn Unsworth at Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia and is used with permission.

The role of the researcher

When meeting the researchers on site or online, at observations or interviews, we came in close contact. In accordance with a constructivist viewpoint, I understand my role as an active and collaborating part of this study. Therefore, some thought has to be devoted to the preconceptions and assumptions I may have brought into this investigation. In other words, an investigator has to ask her/himself what stories are triggered when s/he appears in the setting,

in this case the research group (Denzin & Lincoln, 2008). The interview situation is in the end a power relation between interviewer and respondent (Merriam & Tisdell, 2015) that may be potentially asymmetric in that it favours the part who poses the questions. However, in this particular case, I was a PhD student and many of my interviewees were senior academics, which also has implications for the balance of authority and power. In this context, my profession as a university librarian and subject background are factors that occasionally may have had implications.

Having worked as a university librarian for many years at several different Swedish research libraries, I am used to meeting and assisting researchers in various questions, for instance regarding article publishing or depositing research results in repositories. In my work as a librarian at a university with an open access policy, one work task was also to promote publishing results in open access journals. I suggested solid open access journals to researchers on their request, or more generally informed them of open access publishing as an alternative way of disseminating research results.

Initially, questions about how researchers share their data or not, and the use of the data management plan, might have raised reactions among some researchers. It is possible that they, at an early stage in interviews or observations, interpreted or associated these questions and interest with some form of evaluation or control. If so, researchers might have felt an expectation to make a good impression by for example presenting themselves as being more positive to the idea of sharing data openly than they really were, or by answering in what they believed to be an accommodating manner. My impression is however that the fact that the research group was informed of who I was and what I was doing among them on several occasions balanced the potential bias. Also, I communicated clearly in all interviews that I saw no right, wrong, good, or bad in relation to data management. After having talked with me for a while, researchers in general appeared to comprehend my research interest and my role, and relaxed. In addition, the considerations and scepticism the researchers did express in relation to the data policy indicates that they did not censor themselves. On the contrary, I in fact noted a gradually awakening interest and curiosity on their parts about my perspective on data, particularly among the younger researchers. During the last observation, several made spontaneous contact to talk about data and share their reflections over data issues.

It can be argued whether it is preferable to have advanced knowledge of a world that is to be investigated, or not. With a background in the humanities, I knew little of the details around how natural scientists deal with data on a day-to-day basis. How researchers within these disciplines collaborate or communicate specifically, and what methods and instruments

they used were unknown to me, all factors that made it additionally time-consuming to "conquer" the material. I struggled to orient myself in and understand the environment, terminology, implicit rules, and routines related to the case in order to bring out the essentials for this particular investigation. It remains possible that I have not altogether completely grasped how things are interconnected. I chose instead to see this precondition as an advantage, taking the view that my lack of knowledge of this unknown social world obliged me to be most attentive. I might have noticed things that would have escaped a more familiar eye.

4. Results

This chapter presents the results of the investigation. These results are based on an analysis of the collected empirical material, observations, interviews, and documents. The content is organised according to an illustration of a research data life cycle (see Figure 3 presented in Methods and materials) and is divided in three separate sections, each with a primary analysis addressing the most significant aspects of the case study's results in relation to the research questions. The following sections will frequently report direct quotes from interviews or other types of documents so that the readers may directly partake of the collected data. Because these quotes are transcribed here as faithfully as possible from the interviews, some deviations from standard written English may occur. Emphases and exclamations are highlighted in italics.

Planning for research data

This first section presents the steps of action that the research group took when encountering and interpreting the demands of the funder during the initial project phase, and describes the group's activities when negotiating the data policy, principally during the first six months of the project period; e.g., how they got an overview of project data, how the main communication tool was chosen, how the data policy was introduced to the group and how members were involved in developing the data management plan (DMP).

Before presenting how the research group proceeded with developing the DMP, here is a chronological overview of the project:

- September 2016: Grant application, general strategies for data management were formulated.
- January 2017: Project start, development of the DMP and first contact with library services.
- February 2017: Data policy and preliminary DMP were presented and kick-off meeting was held.
- June 2017: DMP delivered to funding agency.
  March 2018: First project review, no comments received on the delivered DMP

Starting up: orientation and responsibility

At the start, none of the project researchers in the group had had to follow the requirements of a data policy or had written a data management plan previously – this experience was new to

everyone. The co-coordinator said that the research group only had experiences of research funders' requirements about making published articles open access:

> There may have been some rules for publishing, that we had to make the
> publications available by open access at some point, but *definitively* not as strict
> as it is now (C1).

According to the project manager, the funder had declared in the application instructions that writing a DMP would be an obligatory task that should be produced as one of the deliverables. The manager explained that the research group therefore had decided to add a general strategy for how they would develop the future data management plan already when applying for funds. A project member that participated in writing the research application said that adding information about how the project would handle and share data "increases the chances of getting funds, and that's just a reality" (A). The project manager stated that the work with the actual plan began shortly after the project start in January 2017.

At the start, it was unclear who should be responsible for organising the development of the data management plan. The project manager expected it would become her/his responsibility:

> I actually knew right from the start, 'yes, the researchers will manage their part
> but this will end up being my task'. If the project will be granted [funding] I will
> be responsible for it [the data management plan]/…/at least to drive what the
> researchers have handed in. (E1)

In the DMP, the project manager is pointed out as the person who "will oversee data management". The project co-coordinator pointed out that s/he and the project manager were the persons that mainly developed the data management plan with help of some researchers of the group, "no one else felt responsible", they commented (C1).

The project manager explained that in order to understand what data management was, s/he started to look around and contacted researchers at another faculty who had previously formulated DMPs, in order to learn about how the other group had proceeded. The manager also searched online for examples of data management plans from other countries that had long been using them. Finally, the project manager contacted another project officer, in this case the funder, for help:

> I asked their project officer if [s/he] could give me an example of a good data
> management plan. I still thought that it was possible to copy from somewhere
> [smiles], but that did not work… (E1)

The project manager also asked the project officer of the funder for more information on how to write a DMP, but the manager felt s/he did not get much support except for the information that "they [the funder] would not force us to make everything publicly accessible, just the information that is not IPR [intellectual property rights]-sensitive" (E1). After having gathered this information, the project manager said that s/he realised that "data management plans are individual and have to be written by every single project itself" (E1).

An overview of the data

According to the project manager, two documents played an important role in the process of developing the data management plan, the Data Structure and the Data Inventory. After the orienting initial phase and after realising that DMPs needed to be tailored specifically for each project, the project manager felt that "we had to start with drawing a general structure to clarify to ourselves how we should do" (E1). In order to get an overview of the project's data, the manager explained that the group had drawn an image called Data Structure (see Figure 4). This structure sorted out the different types of information that the project generates and which channels are used for these information types, for internal communication as well as external. The figure also showed with arrows how the information types were interconnected and how information flowed between them.

The box called *OwnCloud* (the group's file server solution) that includes the data produced and used within the project and daily work information is central, and all the information flows start from there. *OwnCloud* was where the *internal data* were stored and shared for daily work. Information leaving *OwnCloud* is specified as *patents* in one box and *reports* and *deliverables* in another. The latter are in turn stored in the *Participant portal* which is the project's portal at the funder. The file sharing solution content also has *publications in journals* as another box. The manuscript versions of the publications are stored in the partners' own repositories. The data related to the publications are kept in the data repository Zenodo together with data that were included in applications to the innovation award announced on the project website. *OwnCloud* is described as *confidential*, that is, it is only used by project members.

"We had this [the Data Structure] in our heads, when writing the plan" (E1), the project manager said, and further commented that the project would have had most parts of the structure anyway even without the funders' requirements about data; the solution for file

sharing and the patents were examples of such information and channels. Most parts of this structure were thus described as "obvious" (E1), except the part with open data, which concerned the data that were to be made openly accessible. The project manager pointed to the boxes *Data repository on Zenodo, Data from awards, Open data from research* as the new and uncertain part of the structure and described these as "a question mark" (E1). At this early point in the project, the choice of Zenodo as the repository was preliminary, the project manager said, and could be changed later if they found more suitable repositories when the first samples of data were to be shared. The project manager said that the main part of the boxes included in "Data structure" were general, meaning that each partner would in parallel have their own organisation for their local information.
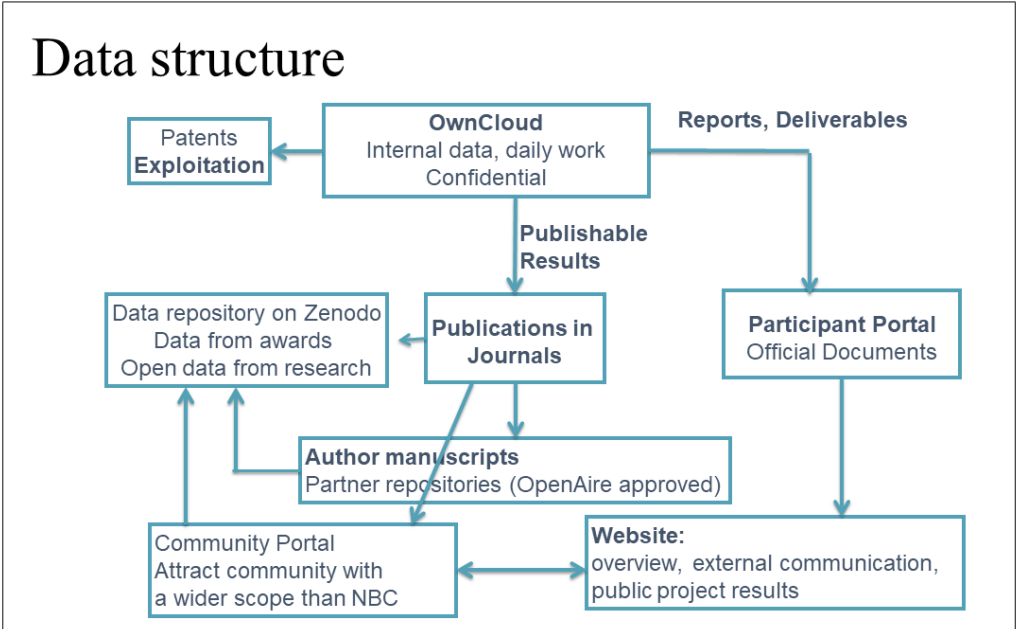
Figure 4. The project document Data Structure. The illustration sorts out the different types of information that the project generated together with the channels that were used for the various information types, for internal as well as external communication. The OwnCloud *box, central in the structure, is the name of the project's file sharing solution. NBC is an acronym for the research area.*

The second document that the project manager said played a central role was Data Inventory (Figure 5), which was the result of a collaboration with a doctoral student to make an inventory of the project's research data. This document was created as the work progressed and did not follow any external template. The project manager explained they had sorted through every project work package and task in order to find what data outputs would be generated. The document lists each task followed by three categories: the task's expected *outcome*, the task's *dataset*, and the *formats, metadata* and *repository* for storing the data. In

some cases, it is specified whether the data types would be public or confidential and several specifications were followed by question marks. The table is difficult for a person without knowledge of the specific tasks to interpret, but it is telling that the group separates outcome from datasets and metadata. Only the latter is used in the group's data management plan. In the first listed task (Task 1.1), outcome and dataset are the same, but in most tasks they differ. For Task 1.3, no outcome is specified but three datasets are declared. Publications and reports are included both as outcomes and datasets. The many question marks in the category format/metadata/repository shows there are uncertainties regarding formats, repositories and whether the material is to be made public or remain confidential.

The project manager summarised the work with the inventory and drawing up a data structure like this:

> It has been *extremely* useful for this project and it wouldn't have gone so well if
> it had not been clear to us how we work with this. So for us *internally* it has
> been useful to draw this structure and clarify how we do. (E1)

For developing the data management plan, the project manager explained that the group had made use of the supporting guidelines for managing data at the European Commission website "H2020 Online Manual" which they had found via the original application template. After the inventory, a first draft of a data management plan was created by the project co-coordinator and the doctoral student using the online template DMPonline provided by the Digital Curation Center.[13] The project's final data management is structured down to the details according to the DMP template for H2020 research projects.[14] The co-coordinator summed up the work with the plan like this:

> We really made the plan very *specific* to what we were doing anyway. In the
> end we mainly wrote down how we are doing this, how we have been doing in
> the past and how we plan to do it in the future. But the nice thing was… I mean
> some things I don't think we would have thought about if we hadn't had to make
> this data management plan, for example this open data (C1).

---

[13] DMPonline maintains templates for data management plans of funders and provide guidance for creating and sharing data management plans, https://dmponline.dcc.ac.uk/
[14] "H2020 templates: Data management plan v 1.0 – 13.10.2016", https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

In line with this statement, a senior researcher added that the goal when writing the plan had been to not change their current routines, "when we were writing this [the data management plan], we started from what we are doing and tried to change that as little as possible" (D1).

Work to be done	Outcome	Dataset	Format /metadata / repository
Task 1.1	algorithm	algorithm	
		Si chips with nanochannels	
		Flourescent microscopy measurements	Mov-files non-public?
		Publication	Pdf, University library repositories
Task 1.2	Simulation software D1.2 and D1.5	Beta version is public	Repository?? Sourceforge?
Task1.3		Si chips with nanochannels	Design-files .csf format (EBL; Raith) or .gds format Search for place for public EBL files
		Flourescent microscopy measurements	Mov-files non-public?
		Publication	Pdf, University library repositories
Task 1.4			
Task 1.5	D 1.4 Report - public		Pdf (cost-free repository)
Task 1.6			
Task 2.1	Optimized junctions		Mov. files
			Design-files .csf format (EBL; Raith) or .gds format Search for place for public EBL files
			Design-files patentable (confidential)??
	simulation software		Matlab? Confidential?
Task 2.2	Programmable junctions/gates (temp sensistive)	Si chips with nanochannels	Synthesis protocols Confidential? Format?
Task 2.3	Filament detectors	Si-chip with nanochannels plus FETs	Mov files, electrical measurements Format? Confidential?
D2.1			Confidential
D2.2			Confidential
D2.3			Confidential
D1.1, D1.2, D1.3			Confidential
Task 3.1	D3.1, D3.2 D3.3	Reports	Confidential
Task 3.2			Synthesis protocol (patentable; CO)
Task 3.3			
Task 4.1	Exact cover devices D4.1	public	Design-files .csf format (EBL; Raith) or .gds format Search for place for public EBL files
Task 4.2	D4.2 public		.mov files
Task 4.3	D4.3 confidential		
WP 5			
Task 5.1	Stakeholder workshops	Roadmap of challenges	Pdf- file (cost-free repository)
Task 5.2	Training periods		No-data available
Task 5.3	Open innovation awards	Public domain	Pdf –files (cost-free repository)
Website and community portal			[Coordinating university] is hosting
Dissemination	Research publications		Pdf (cost-free repositories)
	Youtube clip		Social media
	Logotype		
IP protection			Patent databases

Figure 5. The project document Data Inventory. The project group has specified each task's expected outcome, dataset and format, metadata and repository for storage. (Data Management Plan)

Introducing the data policy to the research group

A few of the researchers had encountered the data policy when writing the research application. However, the project manager and the co-coordinator formally introduced the funder's requirements about data and data management to the whole project group at a project kick-off meeting in early February 2017, one month after project start. In presentation slides, they showed the Data Structure and explained that a data management plan defines issues such as how data will be created and documented, who access them, where they will be stored and whether and how data will be shared. One slide described advantages of data management: "well-managed data opens up opportunities for reuse, sharing and makes for better science!" Data management was said to make research easier, that researchers could avoid accusations of fraud or bad science, and in the end, they would get credit for sharing data. At the bottom of the same slide, it was stated that the project should meet the funders' requirements "with minimal effort". At the meeting, the presenters asked all researchers to complete the draft of the DMP. The group members had, according to another slide, to "think of the output of each of your tasks and answer the guiding questions" and send input to the project manager within shortly.

The project manager said that s/he remembered that the project researchers initially reacted to the presentations of the funder's demands with grins and sighs, and expressed worries over increased work effort and bureaucracy. These feelings were reflected also in interviews where a researcher complained "it is too much of documents today" (D1). The project manager said that at the meeting, the group did not see any value in developing a data management plan or sharing their data openly. One researcher said in an interview that to sit down and describe how they were working was "a waste of time, /…/ that's always the case whenever you have to document things when there's no direct benefit" (A). The project manager said that s/he thought having a DMP and sharing data openly would be valuable, but that it would take years before that value would show; that is, before other researchers would be able to make use of this groups' data and publish new results.

A couple of researchers thought that the task of describing how they managed data was unnecessary, since these activities were obvious parts of a scientist's work:

> I really hate these plans here and plans there etcetera, it's just a thing you... you
> should do it anyway. If you are a scientist you would know that you keep the
> data like this/…/ You are completely lost if you are not doing the

> documentation, that you can find what you have been doing and where the data
> is (D1).

Other researchers corroborated this perspective, describing that data management was already integrated in their daily work as standard practice, stating:

> This is exactly/…/what we have *in place* in our lab /…/. We store it [the data]
> on the server, we keep it on the server for a certain time with fast access, and
> then we put it in the archive. And this is documented by all the scientists, in
> their lab books, and in their write-ups, so this guarantees that it's always
> findable again and re-usable. (C3)

Another researcher stated that the data management plan did not alter daily work because the group was already thinking of issues like data formats and the receiving parties when they were communicating data:

> [We] exchange data all the time... with so many colleagues, so many people
> and... We have to look at the compatibility and we have some experience of this
> so that we know 'OK this data format can be read out of the public licence so
> we use PDFs or something like this'. But this is what we do all the time, so
> without data management and with data management... [It] does not affect our
> daily work so much (B3).

Nor this person's colleague thought that the plan added extraordinary requirements to the research work, but only made ordinary data management "more formalised" (B1). A third researcher in this group agreed and said in summary the that the data management plan had no real influence on how s/he treated her/his data except that "we have it now in our heads" (B2). Another researcher stated that s/he thought the group carried out experiments and data processing as before they had had the plan, but that the policy now "forces us to make things available to others" (E2).

Several researchers mentioned that functional data management was important because they, like all researchers, had to be prepared to argue or defend their research results. They recounted experiences in their careers when they had received questions or comments on their studies from other researchers, and they had then had to look up their saved data, sometimes as long as a decade after publishing. Therefore, good data management was something that "has to be done as a scientist", one researcher stated (B1). Another researcher described that data management routines solved problems of finding and understanding the data from former colleagues that had left the laboratories, "then it's hopeless when you want to publish

something, to try to find what's been done /.../ Therefore it [the data] has to be so clearly described that you can understand" (D1).

A young researcher expressed that from her/his perspective, as a new researcher who was beginning to develop new data sharing routines, having a data management plan was helpful: "for me it's more like a guideline" (D4). This researcher thought the conventions for file naming and storing facilitated access to data both within the group and for outside researchers, and also over time, "so even after ten years, [when] someone tries to find it [the data], from the date, the name signature and type of experiment, [s/he] can find it /.../ If they want to check, it's not difficult" (D4). This person's colleague reasoned that the usefulness of the DMP might differ between young and senior researchers:

> For somebody who is already at the peak of your career you already have your
> own system. For us starting our careers it's good to have a universal code about
> how to save the files, for everybody to act. It would be good starting PhDs… to
> teach them like how to save data. (D4)

> I think you have a point because this way of storing data and saving data, it's
> something that you learn when you're starting up as a PhD student... And then
> you just do it because you're used to it and then you don't have to have this plan,
> because you already do that, do what [it] says in the plan. (D2)

A couple of researchers mentioned that a data management plan could benefit a context in which people from different countries were collaborating, because then no further discussions about how data should be managed would be needed. One of them explained, "I think that's highly helpful especially as you're not just two groups working together /.../ to have it standardised" (E2). Another researcher thought that having to make data openly accessible could lead to new possibilities, but that there were technical problems with making this project's data accessible:

> [What] can come out of sharing the data is that you have another publication or
> a collaboration… that someone else uses the data for something you didn't even
> think was possible. So in principle I am very *for* sharing data openly. But for our
> data it's very difficult because the files are so huge, I mean, I have terabytes of
> raw data and it's difficult to provide enough web space for that. (C1)

Finally, one researcher stated that making data openly accessible could be "*very* important, but only sometimes" (A). This researcher used an example of the medical field in which s/he

worked in parallel, and where s/he thought data policies could enable increased verification and control. S/he considered the funder's initiative as "noble but didn't really understand the problem" (A). The researcher said that while this project group did not "want to obscure anything" (A), the data policy would have no effect because no one would be interested in the data before they succeeded in solving the research problem; the data management plan "is *not* a part of our lives" (A). S/he added that, at present, people who are interested in the project data already knew of them or "they would just *ask* me if they wanted something from me" (A). The researcher reflected that if the research project succeeded in solving its research problem, this would "...be a break-through of such magnitude that people would get their hands on that data by other means" (A). S/he illustrated this by drawing a parallel to how the conference papers where the founders of Google first published their ideas did not attract any interest until their company had become a success, but then these papers suddenly became highly cited and spread.

The file sharing solution – the key communication tool

When talking about the development of the DMP, both the project manager and the co-coordinator related to the previously mentioned online cloud service they used. This file sharing solution enables online collaboration and storage, and is open-source software that is hosted on a home server that does not offer storage capacity. In this project, this service is hosted by one of the partners' universities. According to the researchers, they use the service for storing and sharing different informational material including reports, meeting minutes, agreements, patents, press releases, agendas, research data and the data management plan. I observed how it was used at a meeting where all members were instructed to add information to a document put on the file sharing solution before the next day. All researchers had access and editing rights to the file sharing solution. In several interviews, the researchers said they expected this tool to be of use when members of different partner groups wrote publications together.

The project manager described the file sharing solution as central when explaining the Data Structure picture: "the core is [the file sharing solution] where everybody works together" (E1). The file sharing solution is mentioned a total of four times in the DMP in descriptions of how data will be shared within the project between project partners. When the overview of the project data had been formed, the co-coordinator had presented the idea of

using the file sharing solution at the kick-off meeting as a "central data hub where we can share the data and that sort of stuff, and around that we built our data management plan" (C1). The co-coordinator further explained that exchanging large data files with project members had been difficult in previous project periods and that the new file sharing solution had made the exchange easier; facilitating internal data sharing had brought group members closer together and also resolved preoccupations with securing the data:

> I feel this was a big step forward to /.../shrink the internationality. From a data perspective, to be able to share data across borders, via the internet. This just wasn't as easy a couple of years ago. And even in [the preceding project] we didn't even have Dropbox, that just didn't exist back then and there it was really, really difficult. We could only share evaluated data via email. Large files we would have to either bring during project meetings or come and visit, so that was very tricky/.../And now with [the file sharing solution], part of because it's hosted by [a project partner's university] we don't have any privacy concerns /.../ [The previous file sharing solution] does the same thing but /.../ basically it ends up on some international US server and this could create concerns with patenting and IP [intellectual property] protection /.../ So now with [the file sharing solution] it really feels, *to me*, for the first time really effortless to share across countries. (C1)

The researchers in the group were all instructed to import the file sharing solution to their computers' local file structures in order to have the system synced and easily accessible. To conclude, the co-coordinator stated the file sharing solution had definitely changed how the project dealt with data: "it's much easier to send files back and forth and everyone always has the key information on their hard disk at any time so that really is very helpful" (C1). Additionally, the project manager described the significance of the sharing solution and data sharing:

> I think this online collaboration – to *have* your data and to *share* your data on this [the file sharing solution] – is enormously valuable for keeping this project together. (E1)

Mediator bridging between disciplines

The project manager told that after the first draft of a data management plan was made, the co-coordinator had sent it to a few researchers that s/he knew had the right knowledge (about

for instance data types and how data could be made openly accessible) to complete it. The co-coordinator described how the whole group after that had been involved and the approach of this process further:

> C1: So actually the *main work* was really just *writing down* what we knew we
> wanted to do anyway and then also to get feedback from others, 'OK how are
> you doing it? Do you want to add something?' So we sent it around to everyone,
> and they… some people added a few lines here and there, or also they just said
> 'oh yes that's fine'/.../ we just consulted everyone, like 'is it OK for you?'
>
> Madeleine: So that everyone had the opportunity to object?
>
> C1: Exactly. But this was the main reason for not putting to many rules in there,
> because we didn't want to impose on people and force them to do something.
> We just felt it's nicer to sort of show them 'OK these opportunities exist' rather
> than trying to force them do something.

After having received feedback from the group, the project co-coordinator made final adjustments to the data management plan. The project manager described her/himself as the right person to do this, since s/he had a background in a different field in which data sharing was common. For this reason, the project manager said, s/he had insights about how data were handled in different disciplines, for example, which data repositories were commonly preferred. Another researcher mentioned that the broad competence of the co-coordinator gave her/him a mediating role, linking between group members:

> Sometimes we have one person working on one end and another person on that
> end... *Ideally* they can talk to each other, but sometimes it needs a person in
> between, that can communicate between them /…/ [The co-coordinator] who is
> working with us is a very good person because [s/he] understands a little bit of
> this and a little bit of that so [the co-coordinator] can sometimes try to translate
> for us. (C3)

Data-related issues in need of library support before plan delivery

To meet the requirements of the data policy, the project needed external help to deal with several issues. The project manager said s/he therefore contacted the faculty librarian who put her/him in contact with librarians working with data management support, of whom one (G) was interviewed for this study. In an email dated January 2017, s/he posed questions about,

for instance, whether the data management plan had to be written from scratch and how they were going to manage data that were neither PDF files nor publications. When the project manager met with the librarians a month later, the librarian (G) said that s/he and the project manager had discussed what could be considered a reasonable level of detail for the DMP and how the project could avoid spending too much time on it. They had also discussed the issue of whether the funder would just check off the data management plan as a submitted deliverable or give the project detailed feedback at the forthcoming yearly review.[15] The librarian (G) stated that the project manager had expressed concern at having to burden the project researchers with additional and time-consuming tasks related to the DMP and making data openly accessible. The project manager had felt it was her/his task to motivate the group to contribute and reach consensus under the difficult conditions of considerable geographical separation and across different disciplines.

Regarding the question of metadata, the librarian said s/he and the librarian (G) had not found any standard vocabulary that fit the interdisciplinary character of the project, which had been noted in the data management plan (this issue is further commented upon in the upcoming section "Methods for making data accessible and interoperable"). The librarian said that s/he and the project manager had discussed metadata and the problem of finding a controlled vocabulary suiting *both* the project's nascent topic *as well as* its interdisciplinary character:

> One has different ways of working with data and talk of data and to assess how one does and the difficulties with that… And when one should reach consensus on this… like 'how do we manage data so that *others* understand what we have collected and analysed this and processed it?' (G)

For the choice of data repository, the librarian said s/he had recommended that the project manager consider the target groups for the project's data and the pros and cons with choosing a specialised versus a more generalised repository. The librarian and her/his colleagues had also informed the project manager of the risk of future added costs of choosing a commercial repository.

The librarian and the project manager had further email contact two months later, in April 2017. In these emails the project manager asked for help with concretising the issues that concerned keywords, DOIs (Digital Objective Identifiers), and FAIR (Findable, Accessible,

---

[15] At project reviews, the research group present achieved research results and forthcoming planned work to project officers and external reviewers. In this way, the funder monitored the progress of the research project as it proceeded. This research group had three reviews planned during the five-year project period.

Interoperable and Re-usable) data in the DMP. The interviewed librarian recalled having had "stressed the issue of discoverability and re-usability" (G) when commenting on the plan in order to explain the purpose of adding metadata.

The project group received no feedback on the final delivered DMP at the first project review in March 2018. The co-coordinator expressed disappointment over this matter saying, "we put a lot of thought into it" (C1). The original DMP had not been updated or changed as per May 2021, one year before the end of the project.

Analytical summary

To begin with, it is the community's shared interest in developing the project's research that makes project data be produced. The Data Structure document's description of the project's information flows and channels tells of how the group is engaged in producing and communicating data of different forms. The document shows that the group worked in common to shape the organisation and communication of their data and the ways in which they agreed to work with the data to reach their common goal. We see that the group included many different types of information as data, such as reports, internal data, patents, and publications. The group used shared terminology such as *data*, *outcome,* and *datasets* to describe what was produced. For an outsider, it is not evident what these items include or how they differ, but it appears to be apparent to the project members who are used to these terms.

In the material, we see a number of activities that the researchers carried out as a response to the funder's requirements and data policy, which were new experiences for everyone. None of the researchers mentioned the option of opting out of the Open Research Data Pilot or if this possibility had been considered. The decision was to comply with the policy that was perceived to be an "obligatory" deliverable (E1). Forming a data management strategy early in the application stage was also motivated by the idea of gaining potential advantages and to enhance the possibilities of receiving research funding, as (A) stated. The community worked to both deliver a data management plan in time and find ways to interpret and meet the expectations, specifically during the first months of the project period. At first, three persons in the group (not the whole group) were mainly involved in interpreting the data policy: the co-coordinator, the project manager, and a doctoral student. The first two were part of the management team that decided to take the responsibility for leading the work so that a DMP would be delivered. They interpreted the demands and shaped the immediate local response.

To shape a response to the data policy required efforts and time from many researchers, initially mostly by management. Management initiated activities like trying to understand the policy, gathering information, contacting persons for support, forming an overview of all project data that were to be generated during the funding period, informing, and engaging all community members, and motivating them to contribute their specific competence to the final plan before delivery. The funder's directives were interpreted by management as vague, the support was seen as insufficient, and management had above all little time to make use of existing information. During this intense phase, management learned what a 'data management plan' was, that it had to be tailored for this specific project's data and activities, and what information they needed before being able to fill in the DMP templates.

Confronting this work made it evident that necessary knowledge was lacking within the group, and competence had to be looked for in parties outside of the community. Communication between management and the librarians describe to some extent what competence the community lacked if they were to be able to meet the policy. The concerns included several issues of a more technical character such as metadata standards, interoperable identifiers, available data repositories, and how to make project data FAIR in practice. The outside helpers, to varying degrees, came to serve the community as interpreters of the funder's requirements. In clarifying and explaining things to the project manager, they bridged the community and the funding agency during the early stages when the group was struggling with responding to the demands. New knowledge and tools enabled the group to move forward in the work of developing the data management plan.

The detailed descriptions created to overview the project research data and the new documents Data Structure and Data Inventory became new common tools for the project. These served to explicitly frame and summarise all details of the data produced by the six project partners in a new way regarding the data formats, appropriate repositories, and the community's information flow between external and internal channels. The project manager's description of these documents as "*extremely* useful /…/ for us *internally*" (E1), stating they clarified how the group could continue the work, indicated that reviewing all data in detail was important when responding to the data policy.

The file sharing solution, the community's prominent communication tool for internal data and information sharing, is another tool that the results show played a key role in the researchers' work of responding to the data policy. This sharing solution constituted part of their shared repertoire and is where much ongoing collaboration takes place, "where everybody works together" (C1). The sharing solution functioned as an information node that

connected the community because it could be aligned with the community's needs. In enabling sharing of the community's data, even with the large data files that they had previously had problems and in bringing together community members who were far apart geographically, the tool contributed to sustaining dense relations between community participants. Sharing data using this tool was described as "effortless" (C1). Integrating the tool was one of the management's deliberate strategies for enabling data and information sharing among participants. The co-coordinator's statement that the data management plan was planned around this tool (C1) implies that the ideas for organising and communicating the project's data within the group were closely interwoven with the overarching social organisation in general.

The material shows that when the data policy entered the work of all group members broadly at the kick-off meeting, the discourse had already been shaped by management to a large degree. Management's positive approach to data sharing ("I am very *for* sharing data openly", C1) is notable and suggests that the discourse was actively guided or shaped. The information in the meeting's slides showed how the policy was to be approached by the community, that management had assessed that the appropriate action, and that the mutual approach was to comply to the data policy with "minimal effort". Management also made efforts to encourage and motivate the community, explaining the advantages of data management and not wanting to "impose on people and force them" (C1) or take up their time. The "*main work*" was to write down what they had decided to do, as the co-coordinator said, and by doing this, managements work saved time for the other participants. Even if the overall discourse was shaped, all community participants were heard and asked to take action; they could complete, accommodate, or adjust the preliminary DMP to make it suit their specific data and activities. Management asking all researchers to contribute to the data management plan draft ("we just consulted everyone, like 'is it OK for you?'", C1) shows that management tried to reassure that everyone was on board.

The participant's statements show a variety of opinions about whether the data policy was understood as meaningful or not. The sighs and use of strong words like "forced" (E2) shows that some researchers experienced a strong sense of compulsion and constraint in relation to the data policy. Fear of increased bureaucracy consuming valuable research time also formed resistance to the policy. Other expressions such as trying to "change that as little as possible" (D1) and "we mainly wrote down how we are doing" (C1), both relating to the development of the DMP, strongly suggest that the approach was to not change existing practices.

Despite the strategy to reduce work with the data policy, the results however also show efforts taken to comply with the policy and attempts to actually change practices. An example is the inclusion of the data repository in Data Structure, showing the community's intentions to make data openly accessible in the repository. Making data accessible in this way was not an existing practice: the repository was new and the question mark by it in the document indicates that it was *not* "obvious" (E1) like the other information channels were. However, making data accessible was merged into the existing information flows, which reflects the willingness and decision to introduce this activity. The data repository was added as a channel for sharing data as a direct response to the data policy. Sharing data in data repositories systematically had previously not been a part of the researchers' data sharing activities.

Several researchers did not think the data policy was meaningful and used statements like a "waste of time" (A) to describe it. The understanding that research data would be shared instantly if anyone asked for them suggests that the researchers were used to a more informal and research-driven exchange, happening between colleagues that needed the data, on request. The same researcher stated that data that had been used to solve problems, make new discoveries, or verify results were valuable and relevant to share widely, as potentially were data from fields other than the project's. Data with a high level of interest would be accessible in some way independently of data policies. The inability to see any value of using the DMP in the near future, and one researcher's phrase that "there's no direct benefit" (A), were also related to this view. Others did not see what value the data policy could add, as they already considered to have the essential knowledge of data management, which was a known and present activity, an evident part of daily work, carried out "all the time... with so many" (B3)', a part of the research profession that "has to be done as a scientist" (B1). Data sharing did not depend on or need external demands.

Other researchers however experienced the data management plan, produced as a result of the data policy, as meaningful. The junior researcher described the DMP as a "guideline" (D4) that explained how data management should be carried out, which shows that this information helped her/him to understand how the project had agreed to deal with data issues. This researcher and her/his colleague both considered data management to be an activity that was learned in the beginning of the research carrier. This formalisation of data management activities included specific activities like data formats, forms of metadata, and how storage was to be carried out, and was seen as a "universal code" (D4). The data management plan explicitly articulated knowledge that possibly had been tacit and not articulated. The DMP was also considered efficient because it formed a bridge between diverse partners with

potentially diverse practices in collaborations by settling how data management issues were to be carried out. Some researchers appreciated the funding agency's ideas of research data even if adhering to those ideas would use time, saying that sharing could lead to new discoveries and collaborations.

In the material from this early project phase, one can observe how the community planned and expected to share data in an open repository, Zenodo. As a result of how the group negotiated a response to the data policy, their intention was thus to share data via a new channel which they had not used before and integrate this channel as a new practice (this issue will be further discussed in the third section of this chapter, "Data publishing, preservation, and reuse"). The researchers' statements indicate that they understood the implications the data policy could have for their data sharing. Statements that the policy made data management "more formalised" (B1) and that the data management plan was now present "in our heads" (B2) suggest some change in the understandings of data sharing and related issues. That these questions were highlighted and articulated by an external party (the funder) might have raised an awareness of these activities, both of data management and sharing. The same is true also for the junior researcher mentioned previously. However, in both cases, beyond these changes in awareness, this learning, little appears to have happened with the researchers' actual intentions to share their data.

Finally, material aspects of some of the produced data showed to have implications for how researchers share data, and the researchers were used to considering and further adapting their activities because of those aspects. Statements show that these aspects complicated the internal sharing of these data and that this was a problem that needed solving. The co-coordinator said that certain project data could not be shared "because the files are so huge" (C1), meaning sizes in terabytes, which directly connected the material condition of data to sharing them openly, and demonstrates that material aspects can hinder sharing. This idea will be followed-up further in the second section of this chapter, "Generating, processing and analysing data".

Generating, processing, and analysing data

The project data

In this research project, the researchers created their own data. They did not reuse data from others as they developed a new research field, "we are probably the only ones who do this kind of thing" (E2) as one researcher stated. In one interview, the project co-coordinator summarised the project data as "microscopy data, raw data, evaluated data, then publications, posters, talks and then also source code and patents" (C1). Following the funder's data management template, the group described the data types they generated in their own data management plan (DMP), the purpose of generating these data, their type and format, size, and utility to the group and others. The project's data were categorised into five data types: scientific documents, source code, device layouts, microscopy data, and processing and fabrication data (Figure 6).

Data	Purpose	Type and format of data	Size	Utility
Scientific documents	Project reports, research publications, patent applications	PDF and Word	<1 GB	European Commission, Researchers within the field and related fields
Source code	Software for simulations, software for data analysis	Matlab source code	<1 GB	Researchers within the field and related fields
Device layouts	Input for processing and fabrication	2D and 3D architectural elements	<50 GB	Researchers within the field and related fields
Microscopy data	Scanning electron microscopy images for quality control and monitoring	TIFF images	1-5 TB	Researchers within the project
Processing and fabrication data	Scientific documents, device layouts and microscopy data: process flow plans and parameter sets for fabrication	e.g. GDSII	<100 MB	Researchers within the project

Figure 6. This table presents how the five data types the research project generated were described in the DMP. In addition to the data types, the purpose of the data collection, type and format of data, data size, and utility of data are shown. (Data Management Plan)

As seen in Figure 6, the project's data types differ in how they were generated, their format, and their size. Being unfamiliar with this data, I asked the co-coordinator who confirmed that all project data are digital. In the interviews, most researchers stated they worked as experimentalists with the three data types: device layouts (Figure 7), microscopy data (Figure 8) and processing and fabrication data (Figure 9). All the experimentalist researchers presented visible samples of data. A couple of researchers showed short black and white films that were shot with an electron microscope while they explained what was happening in the film, how proteins were moving in the frame and why (Figure 8), which would fit under the category microscopy data in the DMP. Another researcher emailed me a grey-toned image from an electron microscope representing the processing and fabrication data category in the DMP (Figure 9). These images showed a three-dimensional structure within which the proteins moved, and the researcher explained that the process was carried out as a quality control of the fabrication.

Some researchers showed things they called data that did not easily fit in the described five data types of the data management plan's description of data. One researcher emailed examples of what s/he called data objects which contained documentation of the parameter sets for fabrication of a sample. These parameter sets were Excel files with columns filled with registered values (Figure 10). When visiting one partner group's laboratory, the researchers showed me a small laboratory dish that contained proteins embedded in a gel-like matrix, stating these were data (Figure 11). This sample was *per se* not in a digital form, however, the results the sample showed were documented digitally.
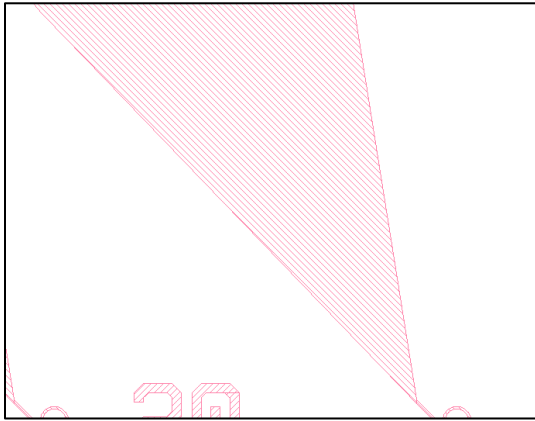
Figure 7. An example of device layout data.
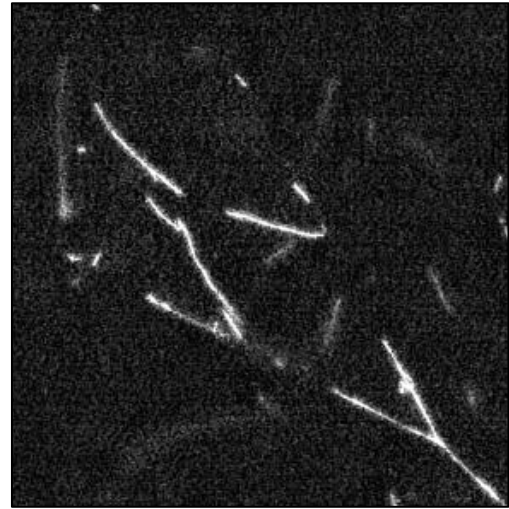
Figure 8. An example of microscopy data, here, a snapshot taken from a film of moving proteins.
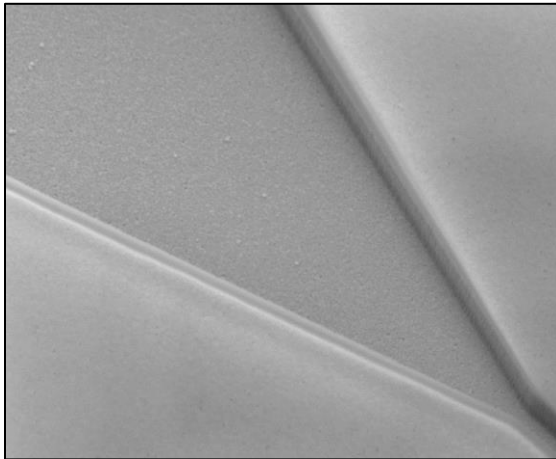
Figure 9. Processing and fabrication data. The electron microscopy image shows a structure in which microtubules move.

Figure 10. A data object documenting the parameter sets for fabrication when a sample was made.
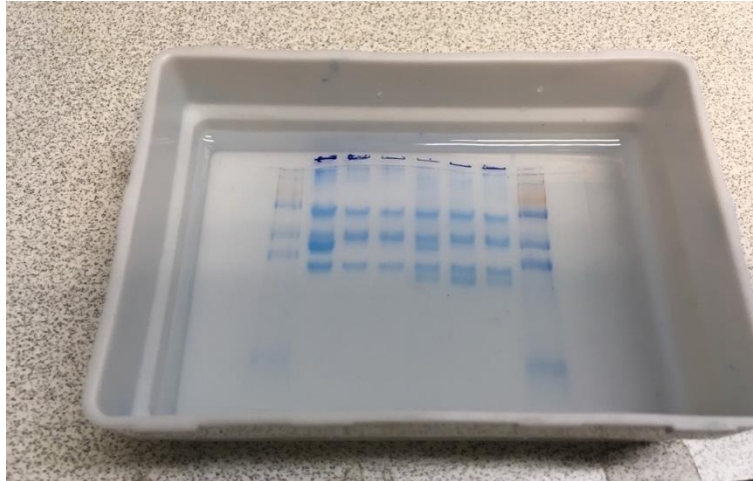
Figure 11. Data in the form of a laboratory dish with proteins in liquid.
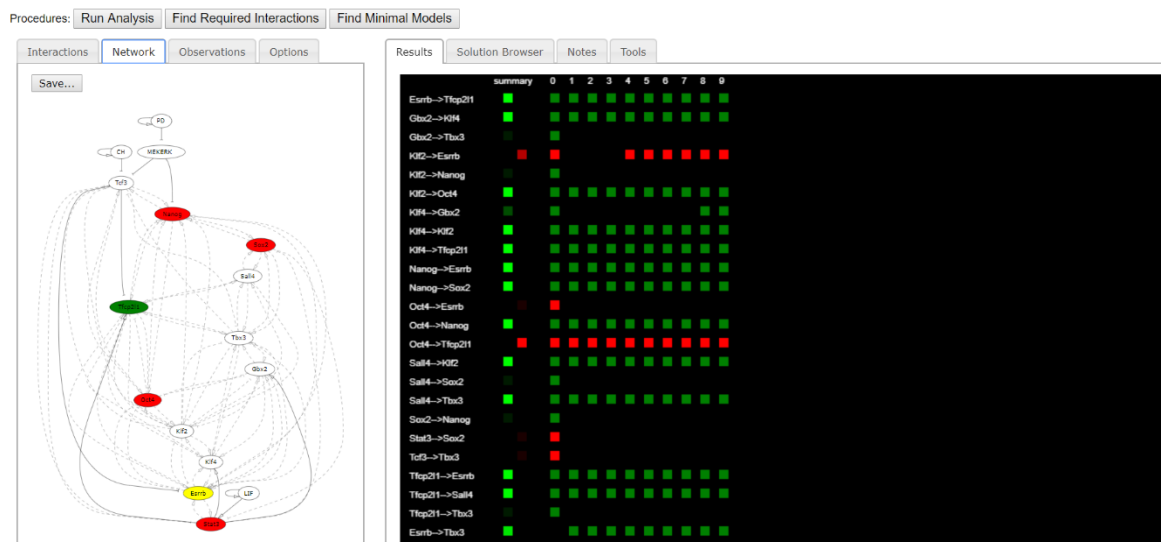
Figure 12. A graphical representation of a network built to process experimental data.

Two of the researchers found it difficult to present visible examples of their data; they both commented they did not generate data in this research project. Their work was more theoretical, and they were occasionally referred to by some of the experimentalist researchers as "the theoreticians". In order to present what they produced, one of them (F, a computer scientist) demonstrated a graphical representation of a network s/he had developed in a different project (Figure 12). The network was similar to the one that F would generate for the present project and was built to process experimental data, and was the product of her/his research work. F's contribution to the project research was "generating the analysis software and thinking about the theory" (F) or doing verifications of experimental data. The researcher said that later in the research process, when a tool or a software that worked would be

developed, what s/he generated would probably fit under the DMP's data category of source code.

A second researcher (A, a mathematician) also could not present anything in order to visualise what s/he generated in the research project. In our interview, A said that s/he felt "shocked" when realising that s/he had "really never thought about what it [data] is" even though s/he "obviously use[s] it all the time" (A). This researcher stated that what s/he generated in the project was "designs", "ideas" or "stories /…/ with lots of mathematics beneath" (A). A also said that her/his generations had "no physical reality" and that any physical example, for example drawing a table of her/his designs, would be "misleading" (A). "It's just like… even *drawing* it is like… sort of sacrificing it a little bit because the drawing isn't… right" (A). In A's opinion, physical representations of her/his ideas could not be sufficient. To the question where these stories or designs were stored, A answered "I keep them in my head mostly… Sometimes I write them down on a piece of paper, but usually just to show someone else" (A). To this researcher, data is what is given when asking a question of nature: "every time I have to ask nature a question, that's data, I think", s/he said, and this data should always have "some degree of certainty". A mentioned the microscopy films made by the other project members and said "I think everything my colleagues produce is data" (A). Because A was thinking about relationships between things instead of asking nature questions, and because these relationships were not certain, A did not produce data, in her/his expressed opinion. What A generated in the project is not included in the DMP.

When asking why the first described data type, the scientific documents, were included as data in the data management plan, the co-coordinator explained in an email that these data offer an instant understanding of the raw data, and that it was a way to comply with the funder's policy for publications:

> In our opinion publications, reports and other documents are a form of processed data. In fact, publications are the most important form of data because they provide a peer reviewed analysis of the results of the experiment that is much quicker to grasp than the raw data.
> Also, the EU has quite strict guidelines when it comes to making publications accessible to the public. Therefore we felt it would be good to cover publications in the data management plan. (C1)

During the interviews, several of the experimentalist researchers talked about different stages of data using terminology typically associated with consecutive stages of research: raw, processed and evaluated data. When demonstrating data, for example a microscopy film or a

still image from a microscopy film, researchers called them *raw data*. After registering these films, researchers opened them in a software that for example could trace certain phenomena. Through this software, data were processed and thus transformed into what the researchers called *processed data*. These data were analysed and further used for example to formulate new algorithms. At the end of this process, results were evaluated, which gave *evaluated data*. The evaluated data had a much smaller format than the raw data that could consist of very large files: "the file that I just opened is 1.5 gigabyte in size" one researcher explained when opening a film during an interview. At observations on site, researchers spoke frequently of these stages of data. In the data management plan, raw data are however only mentioned once.

On the final observation at the research group's workshop, the researchers presented images on slides that showed many of the same or similar pictures as those they had presented previously as data during the interviews. When I asked a newly added researcher in the group (Researcher 1) if s/he thought data had been mentioned or in any other way been present during the group's morning discussions, the researcher answered no. The researcher explained that to her/him, data were the things the microscope produced. One of her/his colleagues, also new to the project (Researcher 2), joined in and said s/he agreed with this definition and that s/he thought what data is, was probably different to me, since what I was working with was ideas. The researcher said that to me (the interviewer), the issue of what data is was a philosophical question, that differed very much from the hard science s/he and the group was representing. This researcher insisted repeatedly that data had to be measurable, that data was what one measures. Therefore, what had been discussed during the day was in her/his view not data but information.

I asked the two experimental researchers if microscopy data that were stored in an archive without any explanatory descriptions attached would still be data. Researcher 1 was amazed by this perspective and answered that they would obviously continue to be data. S/he said that everything the microscope produced was data and that they could never cease to be data. To her/him, the researcher explained, data either were, or they were not; there was no such thing as a state in between.

When mentioning this discussion to the project co-coordinator, this person said s/he thought data had without a doubt been present at the workshop that day. The same researcher explained her/his view of data, what and when data is, and how the term data could be used by examples in an email the same day:

In presentations (PowerPoint or whiteboard), I would definitely consider the following data:

- Anything that represents the results of experiments or simulations, for example:
  - graphs (bar charts, line plots, scatter plots etc.)
  - scientific images such as electron microscopy or fluorescence microscopy (movies)
  - numbers/tables representing the results of experiments or simulations (ideally the numbers are given with error estimations)

- Precise specifications like the layouts used for fabrications

In a more loose sense, I would also consider the following data:
- Estimations of limitations for practical implementation (for example filament mass, run times, energy consumption, etc.)
- Source code
- Mathematical formulations of problems

In the widest sense I would also consider text and drawings that explain concepts as data, such as:
- Schematic drawings of network layouts
- Text that explains the translation of a mathematical problem into network format. (C1)

The researcher ended the email by saying that s/he, as an experimentalist, in day-to-day communication, usually referred to data "as synonymous to experimental results" (C1). S/he added that the theoreticians such as A or F would likely "have different definitions and use of the term" (C1).

Short-term storage and lab books

Continuing with the previously-mentioned example of microscopy data, when asking where data were stored, one partner group automatically saved these raw data on a locally connected computer in the laboratory after they had been generated. This computer was referred to as "the lab PC" by researchers. In the process, when microscopy data were transferred to this local computer, additional metadata, information such as machine settings or exposure time were saved with them. A researcher in this partner group further explained:

Usually it [the data] stays only like for one day on the computer where the microscope is, because we always need to clean it off again, otherwise this computer would get full quickly. (C3)

The researcher's partner group colleague said that backups are made of the data "… to some fileserver which is hosted by our university or our institute and there it stays for a long time" (C1). The analysis of the data is thereby done outside the laboratory because the data are available on every computer via the institutional server. The experimentalist researchers that spoke of storage generally considered these servers to be quite reliable storage places. However, a couple of them mentioned they did backups of data themselves regularly, since they had had previous experiences of losing data. Data from certain machines were saved both at special servers and on external devices, one researcher explained, "...and once a week we do a backup copy on an external device and we keep these separated in case somethings happens" (B5). These researchers knew of the backup frequency of the institutional servers. One of them explained that s/he preferred to trust her/his own security precautions and systematically made a physical copy of the data produced in order to be sure of not losing them in case of fire, flooding, or break-ins (E2). However, in the DMP, data security was a heading that contained only four lines declaring that "secured storage and data recovery" will be managed by Zenodo, GitHub, the group's file sharing solution, and the partners' institutional servers.

Several of the experimental researchers mentioned the use and importance of a "lab book" (see Figure 13) when producing experimental data, but also in other situations. Lab books are paper notebooks in which the researchers collect information about experimental conditions, such as equipment and machine settings. One researcher described them as "a detailed diary of what you have done at work" (D2). When the information of an experiment is saved as digital documents, they are called "lab journals". The lab journal may contain a reference or number indicating a file with the corresponding raw data. At times, the lab journals were printed and pasted in the lab book together with personal comments on the experimental processes and things necessary to remember before the next trial, researcher D2 clarified. Many researchers often referred to the lab journals and lab books as data. At times, data appeared difficult or even impossible to understand or interpret without the lab books.

The experimental researchers in one partner group explained in detail how they used lab books. For instance, they insured that machine data could be stored and saved in a situation when using machines that only had local access and had no connection with institutional servers:

> You write down and then you do it [the experiment] /…/ Or, we write down
> everything we record, that means we don't only rely on the server but we
> protocol every step, what is done physically at the scales or the balance. We also

write it down as a message, so that we later on can consult in the case of some crash on [the] system... then there is a way to exchange data. I think that's pretty convenient. (B1)

This person's partner colleague explained that s/he saved information about experiments digitally for easy accessibility, but also that the colleague gives her/him printed information to follow:

I do have an electronic journal where I put in all my exposures and… or the wafers[16] I use or the parameters so I can easily trace them. Other than that I get like a paper copy, it's like a recipe, which was made either by [colleague] or by [colleague] and I follow this recipe, what they have written down. (B5)

A third partner colleague further clarified the importance of having printed information like this in the laboratory:

We prepare structures to see what kind of process that has to be done or the next step, and this is also data. We have data online at the storage, we have data on our laptops, we have data in the [file sharing solution], but what we need in the lab is data physically on a paper [laughs]. (B3)
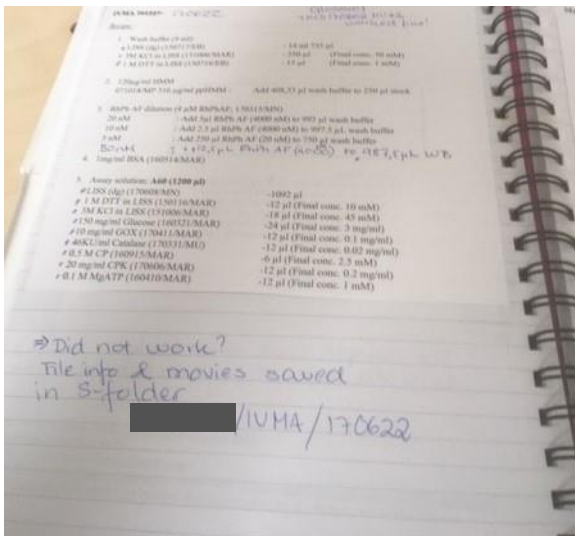
Figure 13. An example of a lab book into which a printed lab journal has been pasted. At the end of the page, written reflections on the experiments and where to find the raw data files have been added.

---

[16] A wafer is a thin oval slice of silicon with microelectronic structures attached.

Lab books were not mentioned in the data management plan, though lab journals were. In the DMP it is stated that the fabrication data that is produced in two individual facilities will be locally saved in in the lab journals.

Processing and selecting data

In the project research, experimental data were in most cases described as being processed immediately after generation. A researcher working with experiments considered that it was easiest if the person who did the experiment and generated the data also did the data processing, since this person had the most information about the process. S/he stated:

> It's the easiest if you process it [the data] yourself because you know what you
> did in the experiment. Of course you have the lab book and you could look it up
> but it's easier to process it [yourself]. (C2)

Accordingly, the person who carries out the experiment often also evaluates the data. Her/his partner colleague explained how the group could work with the entire data process from generation to evaluation:

> The microscopy data in our case it usually *stays with us* because we are the ones
> that also evaluate the data/…/ we actually do the whole pipeline from the
> microscopy data 'til the very end of the useful data ourselves. (C3)

In other cases, the process was divided between the two partner groups. One group would receive data from another group within the project and process their experimental data for them: "they *send movies to us* for data processing /…/ then we do the evaluation, and we send back the evaluated data" (C3). Because microscopy movies are such large files, they were shared via a special branch of the file sharing solution.

This researcher also explained that what is considered to be valuable data were sometimes distilled from initially larger data quantities to avoid the need to share very large files between partners. Data that reflect the present aspect of interest was extracted from the rest of the data, which reduces the size of the files:

> The kind of data that we have, if we should share the movies, that's *lots* of data,
> huge data files, *or* we extract something out of these images and then we just
> send the extracted data which is much less information. /…/ We take a movie,
> but the movie... a lot of the areas in the movies are black. They are not

> interesting because there is nothing happening. So when we *extract* data – like
> the speed of something that moves there – then we have less data but we can
> have all the relevant data. So we can reduce the amount of data a lot by some
> pre-processing. (C3)

In the same way, raw data and process data were at times considered useless, as were non-interpretable data and data without additional information about the experiment:

> The *raw* data, it's not of much use to someone who does not know what I did in
> the experiment so /…/ I mean, that movie that [a colleague] showed you, it
> could also just indicate a failed experiment where there was too much light and
> the microtubules got damaged by the light. So for someone who does not know
> what happened during the experiment, this data is not... it's very difficult to
> interpret. (C1)

Sharing data within the project

In the earliest phases of the project, some of the project partners exchanged few data among them. In December 2017, one partner group stated that project work during this period was local or individual and "each project partner has different tasks so right now we're working towards achieving these tasks" (D3). As was presented in the introduction to this case (see flowchart Figure 2), data were passed back and forth between some partner groups, while between other groups, data were passed on or received without being returned later. The co-coordinator explained that data were being shared between groups for several reasons. For instance, perhaps one partner had access to a specific programme or algorithm that the other was not familiar with, or perhaps this was the only partner who had enough computing resources available to process the data. Other motives were that one partner had more experience for evaluating a specific data type, or to make a colleague that was not involved in the actual experiments evaluate the results to avoid potential bias, or to compare partners' experimental results.

For data exchange between community members, email was used when it was "something quick just for one person" (C3). Data were also shared at meetings or via the common file sharing solution. One researcher said:

> Actually, *most* of the data in a way we share by looking at it at our meetings
> /…/. And for the extracted data we send emails, or we also deposit it on [the file
> sharing solution] that we all have access to. (C3)

The file sharing solution was sometimes used in situations when exchanging larger amounts of selected data with the entire project group: "what we put on the [file sharing solution] is then the *evaluated* data that we want to share with other partners from the [project] team, so... partners outside the lab" (C1). To be able to discuss data at both digital and physical meetings, one researcher explained that files are put on the file sharing solution beforehand, "like while we talk /…/ it's already there, then you go *there* instead of waiting [for it] to upload" (F). This researcher, who said that s/he did not generate data, also pointed out they "do a lot of Skype" (F) to discuss data.

Raw data like microscopy movies were described being too large for communicating via email or the file sharing solution. Further, as mentioned before, these data were in general not considered meaningful to share since they could not be interpreted without explanations:

> The problem is that just for the amount of data we generate here at the
> individual sides is so large. I think it wouldn't make sense to store it in [the file
> sharing solution] because no one would actually know what this data means, it
> must somehow be presented/…/ You always communicate data which has
> already somehow been analysed and interpreted by you. (B2)

Thus, mostly what was called useful data, such as extracted data or results, were what was actually exchanged between partner groups via the file sharing solution, in the form of for example graphs or values. One researcher further explained that this exchange was carried out via the file sharing solution; both parts stored necessary data on the platform, communicating what had been done, enabling each other to check out or comment and thereby continue their research:

> So [a colleague within another partner group] for example stores some excel
> sheets in [the file sharing solution] and then I can go to the Excel sheet and see
> 'ok this is my prepared way for number five' from this charge. And then we can
> discuss about this specific wafer, what's happened with it and I can see 'ok the
> motility that [s/he] wrote is ok' /…/ This is only... an *information* exchange…
> we are just selecting the data we are uploading into the [file sharing solution].
> We don't upload any scanning electron microscope picture in [the file sharing
> solution], *no*, we just select some special pictures /…/ 'Ok' I say 'this is fine',
> and then [s/he] can say 'ok this is... maybe there is a failure in this picture' and

> then [s/he] can say 'ok maybe this is the reason for why motility is not so good
> on this special chip'/…/What we communicate is what we have done. (B3)

This researcher's partner group colleague stated that selecting data before communicating them saves time:

> No one has the time to go through like a thousand images, you just want that…
> usually you have one single question, like 'how big is the channel?' Then you
> want one image where you can see the width of the channel... You don't care
> about the rest. (B4)

Another reason for not sharing raw data via the file sharing solution was that some data were in a proprietary file format. Data that are generated by certain instruments and machines have machine specific formats. Because the partner groups in this research project used different machines, not all researchers had access to the computer programs relevant for opening all generated data formats. "If you don't have that machine, you don't have the program, and you can't… the files are useless for you", the researcher said (B4). To exchange these data with partner colleagues, the data first need to be exported into formats that the recipient can access. Another researcher made a similar statement, "In order to collaborate /…/ we have to export data into interoperable formats" (E1).

    When data were shared between researchers in different disciplines, data appeared to be curated even more carefully than usual, even if there was always an awareness of how to present data when communicating them:

> In very interdisciplinary means... We sometimes need to prepare the data in such
> a way that also the other people can understand it. So that means that data
> preparation and data evaluation need to go a little further step than we would do
> if we were just talking to [colleagues of her/his own discipline]. But at the end
> it's always that you strive to prepare your data in a way that you can present it –
> in a paper, at a conference – to a wider audience. And this we also try to do for
> our internal communication, to present in the same way, and in the same quality
> as we would do for publications, which is also interdisciplinary in a way /…/
> You just *have* to basically *convey* on a level that your audience will be able to
> understand it. (C3)

Another researcher had a similar approach for sharing data: "I'm always thinking about if it's something that's meant to be shared, it should be in a container that's self-sufficient in terms of explanation" (E2). The explanatory information added to the data appeared to increase when the number of recipients was large, "the degree of annotation increases I think with the

level of broadness this data is shared with" (B1). In this researcher's partner group, they spoke in terms of different levels where communication took place, for instance on a data level, machine level, local level, project level or work package level. These levels were thus in part related to the formal work organisation.[17]

Several actions the project took for facilitating data sharing between the partner groups were also mirrored in the data management plan:

> The researchers will curate the data and place data to be shared within the
> consortium in an [the file sharing solution] account held by [the hosting
> university]. The account is personal, and password protected. (Data
> Management Plan)

In the case of the data type device layouts, it is stated in the DMP that a certain format will be "preferred for transfer of information" when two different project partners do fabrication. In addition, part of the device layouts data, and also of the processing and fabrication data, will, according to the plan, both be saved in file formats that "can be opened by open-source software for further modification at the collaboration partners". The project also states that they will consistently follow the conventions regarding file and folder naming described by Stanford Libraries. This convention includes creating a naming scheme that gives information about for example the project name, researcher name, date of experiment, and version number. Except for the junior researcher (D4) that was previously described as thinking that the naming conventions would facilitate access to data, no other researchers mentioned the naming conventions.

Finally, it is worth noting that two partner groups were located in two towns within the same country. As they had the university storage systems in common, they used these systems not only for saving data but also for sharing data.

Data in collaborative analysis

As early as the first observation, the extensive use of images when the group gathered for discussion, presentations, and posters was striking. In subsequent observations, I recognised many of the presentation pictures as the same or similar to those that had been shown to me as samples of data at the interviews. The images presented as slides were tables, still images

---

[17] The research group's work organisation was outlined in chapter three.

from microscopy movies, illustrations, models, calculations, and formulas. These were presented to the group, discussed by the group, and used for explaining or clarifying the research work they had done. The pictures were for example used to share and show the development of an experiment over time. During the observations, phrases like these were common; "here we see… and when we did this we got this…", "when we changed velocity I found out...", and "what happened was that it got too big, it was better with... but we might try...".
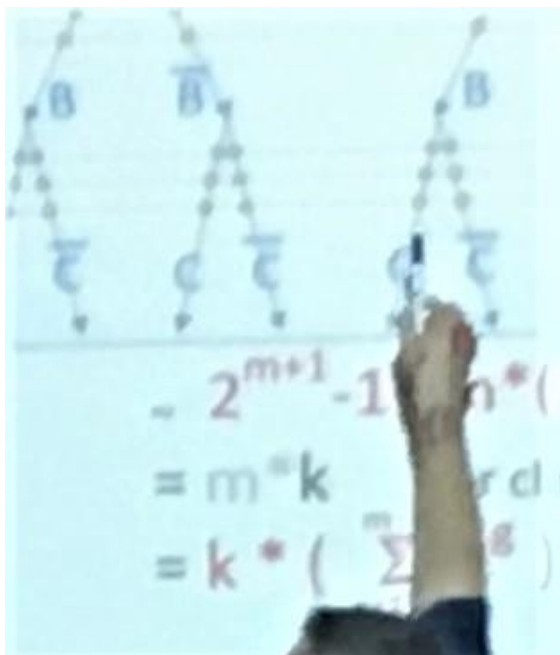
Figure 14. A photograph from a research group meeting where a researcher points at a calculation on a presentation slide.

Researchers also showed images to explain problems with experiments and conclusions that they had drawn from them, "it [the experiment] showed that…we can't get it to stick". Similarly, the listeners used the images when communicating, referred and pointed at them (see Figure 14) when posing questions about methods, processes or courses of action: "why isn't it working?", "is this experimental data or simulation data?", "on what material is this done?", "what temperature did you use?" When the listening researchers didn't understand the presented images, they asked questions about them like "what do the dots mean?" The audience suggested improvements or possible developments to the presenting researchers:

Is there a gap there [pointing at a microscopy image]?

Yes.

What if you'd put on a roof?

Then… [explains what would happen and why this wouldn't work].

During two observations, the fact that the mathematician's outcomes, which were her/his ideas, lacked physical representation appeared to become a problem for the group's collaboration. Both times, the mathematician (A) was asked to visualise her/his ideas or suggestions. The group said that this was necessary for work to move forward. On the first occasion, the mathematician explained her/his ideas drawing on the whiteboard while several researchers commented and discussed the ideas. After that, s/he was requested by several other group members to create a table of her/his idea and distribute it within the group, which would enable the other researchers to have a common reference point for the continuing work, and something formulated that they could test practically by carrying out experiments and fabrications. On the second occasion eight months later, almost the same thing occurred: the mathematician was asked to deliver something visible that explained her/his ideas because otherwise the group could not understand. This situation was more tense, and the researchers argued for some moments. Finally, the co-ordinator expressed why having a visible table was crucial for enabling the research process to move on: "I have a hard time understanding without seeing anything. Are we even talking about the same thing?" (E3).

Analytical summary

This part of the material gives information about the project's data and how a collective mutual account of what data is was formed. The project's common interest in trying out specific new research methods made the data they produced unique: "we are probably the only ones who do this kind of thing" (E2). The intense sharing of data between project members within the community showed that functioning interactions among project members and with many forms of instruments and technology such as electron microscopes, was essential for making most of the data possible.

The co-coordinator's explanation of her/his view on data and use of the term indicates a flexible way of using the term "data", either more strictly or more inclusively. Her/his explanation also implies an understanding that different data can inhabit varying grades of data-ness, from "definitively" being data, to be considered data in a "loose sense" and "data in

the 'widest sense'" (C1). The mathematician's statement that s/he had never thought about what data is, no matter how much s/he had used data and worked with data, suggest that what data is to researchers might be matter that has not been reflected upon.

Researcher 1 and Researcher 2's statements that data is what is measured aligned with the mathematician's understanding that data always needed to have a degree of certainty and the co-coordinator's routine use of data "as synonymous with experimental results" (C1). These views show that they shared a way of addressing what data are. Researcher 1's view of data as what is generated by the microscopes, and that this cannot be anything other than data, tied a close bond between machine and data, and led to an understanding of machines as originators of data or data makers. Because of the machine that generates them, data become data, and never cease to be data. What was data did not depend on whether they were understandable or interpretable. As a consequence, data-ness appeared to be independent of the possibilities of making meaning or sense, and therefore also of being used as evidence. Data will be data even when they cannot be understood, so in this view, usefulness was not seen as an intrinsic quality of data.

The data management plan and the interviews showed that the researchers included a wide variety of materials in the concept of data, for instance project reports, patent applications, microscopy data, and talks. From the perspective of data as information used as evidence for research (the working definition in this thesis) it was surprising that scientific documents were seen as a data type (as they were in the project's DMP). These documents may contain data and are evidently closely related to the project's research processes and results, but it does not seem possible to immediately use them as evidence for research purposes as would for instance the microscopy data. However, what justified the documents as data was in this case the understanding of these materials as "processed data" (C1) and their function of providing fast access to analyses, and explained and reviewed data, which were particularly valid for the research publications. By this connection and function, these documents were justified as being data. In addition, attempts to satisfy the funding agency were also explained as having contributed to the group including scientific documents in the DMP as a separate data type. The funding agency was in this case seen both in the role of an external recipient of the data management plan and as the source of the requirement for the open access policy for publications which the project had to follow. The community negotiated and simultaneously integrated the various requirements of the funder, working with shaping a response to these which thus had effects on what in the end was listed as data in the DMP.

Although it was stated both in interviews and the data management plan that all project data were digital, the lab books and the protein samples were examples of physical objects that were referred to as data, which shows that non-digital data also made up part of the project data. The protein samples evidently constituted actual evidence in themselves, but such material is rarely included in discussions or policies about open access to research data. The data policy at work in this context focused on making the digital data accessible, not physical objects that were data. Lab books were viewed by several researchers as intimately connected to the research data, or as data in and of themselves, as these contained valuable or necessary accompanying information *about* the data. These books were essential tools for enabling a full understanding of the experimental data and functioned as memory records for documentation and assurance of collective information of the experimental processes for future use. Interestingly, even though the lab books were assigned so much importance, they were left out of the DMP, as were the protein samples. The reason might be that they were not digital and could not be made openly accessible, as opposed to the digital lab journals and the other data in the data management plan. If so, it is the material representation of these data that made researchers exclude them from the DMP. The lab books were also stated to be of a personal nature, which suggests another possible motive for not including them.

The researchers' descriptions of themselves and the other participants when presenting their data highlighted a new aspect of the social complexity of the community. Many of the researchers spoke of two groups within the project group, the experimentalists and the theoreticians, as forming two identities. All appeared to be aware of who exactly belonged to each community of practice – no doubts about who belonged where appeared in the researchers' statements, and the understanding of where the boundary was drawn was clear. These articulated identities appeared to be much more prominent than the researchers' more formal disciplinary identities, which were rarely mentioned.

That the participants belonged to different communities of practice came to shape what data they worked with. The co-coordinator commented on the differences between experimentalists and theoreticians, reflecting an awareness of how these memberships in the other communities of practice shaped how the word "data" was used and defined. Community of practice membership also shaped whether they considered what they worked with or generated to be data at all (the experimentalists did, the theoreticians did not). However, the mathematician (A), who called her/himself a theoretician, and the experimentalists agreed in how they defined data: data was something that was measured and certain, and that which comes out of certain machines. The mathematician explicitly stated that this way of defining

data meant that what s/he generated in this research project was excluded from being data; what s/he generated did not fit into this definition's criterion. The theoreticians therefore needed and used other, more relevant words, like "designs", "ideas" and "stories" (A) or "thinking about the theory" or "doing verifications" (F), to describe what they generated within the community. For the experimentalists, the word "data" appeared to be used more directly and self-evidently.

In addition to the use of terminology, the material representation of data differed between the two inherent communities of practice. Most of the experimentalists' data had specific digital formats, some of them characterised particularly by their large size. While the computer scientist (F) produced what would later in the process fit under the category source data and be in a digital format, the mathematician's contributions to the community did not have any physical reality and left no visible traces. No matter how valuable the mathematician's contribution was in driving the idea behind the research project, the lack of visual representation of her/his contributions, her/his lack of willingness to draw them down, on two occasions created tension and confusion. When coming together for a common enterprise, the diversity in the two communities of practice showed. In this matter, they did not agree on what mattered or on the appropriate action. The experimentalists could not carry on their work without visible instructions like tables to follow. In fact, as the coordinator stated, in lacking visible data, the experimentalists could not be certain they were discussing the same *thing*, which disrupted both work and the relations between the researchers. The mathematician on the other hand considered physical representations of her/his data to be misrepresenting and insufficient. Because the mathematician was the only representative of her/his discipline within the project, it is difficult to know whether this researcher's understanding of data and resistance to drawing her/his ideas down, is valid for other mathematicians. An alternative interpretation could be that this behaviour was a result of individual understanding or a situational expression. The fact that similar situations occurred twice however confirms this lack of visual data was problematic for the community.

The material shows many aspects of how researchers share data with each other and the priorities they make when carrying out this activity, which is valuable for understanding the researcher's shaping of the response to the data policy. In the co-coordinator's descriptions of the many motives for why data are shared internally, the intensity of data sharing between partner groups and participants was notable. Competence, technology, result evaluation, and storage space were reasons for sharing data between participants. Data sharing is an activity that results from the community's mutual engagement. These needs links the interdisciplinary

participants together: they are dependent on each other's competence if they are to reach their common goal. The data sharing practice is an activity interwoven in the research practice and its purpose to develop knowledge.

The statements about data sharing tell of common and negotiated ideas about what mattered in this activity, and how it was going to be carried out in order to make collaboration function. Besides sharing almost exclusively selected data, researchers routinely controlled and curated data, adding information about them in order to prepare and contextualise them before exchange. Proximity to data was highly valued within the community, as this closeness provided the best conditions for being able to understand the data, "because you know what you did in the experiment" (C2). Because the diversity of researchers was key in the strive towards the common enterprise and special competence were needed for processing data, the processes often had to involve more than one partner group and data were thus exchanged between these. Communication channels, particularly the file sharing solution, were tools that enabled communications and data sharing, which contributed to keeping the work processes and researchers together, bridging the physical distances between partner groups.

Another issue considered when exchanging data was format, as it was necessary for enabling communication and sustaining relationships: "In order to collaborate /…/ we have to export data into interoperable formats" (E1). The recipient's knowledge was considered and how the data had to be explained to make them understand the data, for instance, "I'm always thinking about if it's something that's meant to be shared, it should be in a container that's self-sufficient in terms of explanation" (E2). Data size, quantity of data, data type, the recipient's discipline, and the number of persons with whom the data should be exchanged were also among the many aspects taken in consideration when sharing data. In general, the statements show that researchers only exchanged data that were well explained, and understandable; this was considered meaningful data sharing.

Statements regarding the valued ways of sharing data described how researchers shared data internally, and this routine way of thinking about useful ways of sharing knowledge did not appear to be confined to only internal sharing. Statements described how data always were prepared in a presentable way, without separating internal sharing from external: "[a]nd this we also try to do for our internal communication, to present in the same way…" (C3). Their approach demonstrates that the researchers explicitly connected their awareness about how data sharing should be carried out within the group to communication and sharing of other kinds or information, in other contexts such as conferences. Data exchange in this way was carried out in line with how other information was communicated. Researchers were

112

interested in communicating in an understandable way, independently of any particular situation.

That the mathematician's data had no physical representation might also explain why these data were not represented in the data management plan. The mathematician stored her/his ideas in her/his mind and only gave the ideas a material representation and wrote them down on paper in situations when s/he wanted to show them to another person. The format of these data cannot be specified; they need no storage place and cannot be made openly accessible without contacting the researcher generating them. Here it is important to not ignore that in this respect, the two theoreticians' data were different from each other. The things the computer scientist generated, or at least the material representation of what her/his work resulted in, were represented in the DMP in the form of source code that were planned to be made openly accessible.

Data publishing, preservation, and reuse

Open data and closed data

In one section of the data management plan (DMP), the project group described how they would proceed in order to follow the FAIR data principles of the funder. The group explained that generally, regarding all project data, "data that are exploitable will be kept confidential. Data that are not exploitable or have been protected will be used to produce scientific, peer-reviewed publications" (Data Management Plan). This explanation makes the more detailed descriptions of which data they plan to make "openly accessible" and which to keep "closed" (following the funding agency's template for the data management plan) overlap, because each data type is planned to be partly protected.

The project group had chosen separate channels for the three data types that in part would be made openly accessible. Scientific texts in the form of community building documents, public project reports, and author manuscripts were declared to be accessible via each partner's institutional repository together with related data. Part of the developed source code data was declared to be made openly accessible on GitHub "after IP protection and/or first scientific publication". The group said the fabrication data was sufficiently outlined both in internal reporting and in research publications to allow reproduction of experiments by fellow researchers, which is why they would not be made accessible elsewhere. These outlines were thus another way of making data accessible.

The data types listed as closed data were followed by explanations for why they would not be shared, as demonstrated by this quote from the DMP:

1. Scientific documents: closed project reports. Reason: relevant for IP (Intellectual Property) protection.
2. Source code of unpublished programs. Reason: quality management. Relevant for IP protection.
3. Device layouts. Reason: highly relevant for IP protection.
4. Microscopy data. Reason: relevant for IP protection.
5. Fabrication data. Reason: relevant for IP protection. (Data Management Plan)

The device layout data and microscopy data appear thus to remain as completely closed data, used only by project members. The results show that data of more or less all five types produced within the project would not be shared or made openly accessible at all, or at least not before the researchers themselves had used these data for publishing scientific articles or exploited them in other ways. The data management plan did not state a time when the group

intended to make data openly accessible; the only time sequence stated was related to the source code that would be made openly accessible *after* they were no longer protected or after a scientific publication.

In addition to the data management plan, the responses to the one-question survey made nearly two years into the project period (November 2018) showed that five of the 15 researchers that answered had made project data openly accessible up to that point. Three of these five researchers specified they had added project data as "supporting information" or "supplementing materials" in journals. The activity of publishing data alongside articles was mentioned also in interviews, as an accepted procedure: "I mean you always have some supplementary for a paper, like some movies" (C2). The other two researchers answered that they had made data accessible on GitHub, one of them commenting that s/he had uploaded software used in a paper after having had a request from a person. Three months later (February 2019), eleven scientific articles were listed as project publications on the project website (ten of them in open access journals, the last being a conference paper). For seven of these publications, data were added as supporting or supplemental information consisting of microscopy films or PDF files freely available on the journals' websites. One article linked to GitHub for free access to the source code used to evaluate the data in the publication. The article's method presentation also stated that "the data that support the findings of this study are available from the corresponding author upon reasonable request." In an informal talk with one of the researchers, s/he told me that adding data to their articles was something they wanted to do even when it was not a requirement (F). Adding data to scientific articles as supplemental information was not an activity mentioned in the data management plan. When asking why, the co-coordinator answered, "we did not want to commit ourselves to a fixed strategy" (C1) and related that concern to her or his worries about finding a suitable repository for the group's data.

Lastly, a search in Zenodo in July 2019 for the project name in several publication titles and project member names, as well as on the Zenodo community site, did not render any results. Three months later, the project manager explained that Zenodo had so far only been used for a couple of contributions from the project's innovation award and for two published articles. In June 2021, in the last months of the project's funding period, I asked the project co-coordinator if Zenodo had been useful for the project's data sharing, for making data openly accessible. The co-coordinator answered that in the end the group had not used Zenodo for making data openly accessible for two reasons. First, the repository did not provide enough server space: "no, unfortunately Zenodo was not useful to us. The limit per

repository of 50 GB was too restrictive for our raw data" (C1). S/he s/he wrote that most repositories for large amounts of microscopy data focus on cell biology which is not closely related to this project's subject. The co-coordinator felt it would therefore not make sense to put data there because they would not be found by people who might use them. Second, the project's data appeared not to fit the repository's topics because they had been deleted by repository administrators: "Other data that we had uploaded was discarded by them. This discouraged us from using this repository more" (C1).

Methods for making data accessible and interoperable

Early in the project, measures for long-time storage of data and data preservation were not an issue mentioned often in the interviews. A few methods to be used for making data interoperable or preserving them were stated in the data management plan for example that the group will use PDF/A formats for the first research data type, the scientific documents, "to ensure the documents' portability across systems in a long-term perspective" (Data Management Plan). Data contributing to PhD theses were planned to be stored for ten years, and the project partner who was the student's advisor (or otherwise responsible for the student) was responsible for this task. The project group also stated they would use the public repository GitHub to ensure long-term re-usability of the software generated from the project's source code data.

In the DMP, the project group said that data, both types used only by the project and types made openly accessible, would be provided with metadata. Generally, the group said that metadata would be created in accordance with DataCite with the mandatory fields Identifier, Creator, Title, Publisher, PublicationYear, and ResourceType. However, in the DMP, the group was described as being "encouraged" (Data Management Plan) to add other recommended fields, like for example subject, contributor, and RelatedIdentifier. In addition, adding subject keywords to data for discoverability was encouraged even though "not one controlled vocabulary exactly matches the scope of the project" (Data Management Plan). Because the research group's topic and approach were new, the researchers intended to solve this problem by using other thesauri that "might cover parts of the subject scope" (Data Management Plan) or, as a first step, by members choosing subject keywords freely.

Metadata would be generated both automatically and manually. For the microscopy data, metadata would be provided with "automatically generated metadata like scale, exposure

time, and light source" (Data Management Plan). These metadata would then be completed by manually generated metadata like creator, title or subject. Also, the data connected to research publications was planned to be manually curated by the creating researchers before they were uploaded to public repositories. According to the DMP, this manual curation was to be done after the research publications have been published in institutional repositories "so that the publication can act as an umbrella for finding und understanding the data" (Data Management Plan). The scientific articles put in an institutional repository were to make underlying data findable by linking to open repositories where the data would be uploaded.

The meaning of making data openly accessible and data reuse

The issue of the potential for reusing the project data for the benefit of the researchers or external persons was much discussed in one partner group interview (group B). One researcher did not appear to see much use in reusing data, as data quickly become outdated:

> Sometimes I review results from several years ago just to get a feeling again
> 'what have we done there? What was possible in that moment?' But when I look
> in the reports, on the protocols, the parameters are entirely different than what
> they are now /…/ Everything changes because we are constantly improving the
> processes. (B2)

When asked if they thought anyone would reuse their data five years from now, a couple of researchers had doubts. Data could be used as a memory record, but they become outdated because of the rapid technological changes:

> I want to say… no! There's no interest in the data, because we're developing the
> technology, we are improving so many things and *change* so many things.
> Maybe as a starting point for a new technology development but I think the raw
> data are used just for documentation, not more. (B3)
>
> I agree with that. (B2)
>
> You said we have to go on five years ahead but if I just go back five years and
> think about what's happened with the data five years ago, we don't use it [the
> data]. We improve so many processes of technology... it's an improvement all
> the time so... old data... [laughs]. (B3)

In the data management plan section about data preservation, the group expressed that project data in general would not be useful to third parties for more than five to ten years after the

project's end. The reason given was that newer data will be more relevant. The rapid development of the tools, the machines and software, also makes data change and, more importantly, the skills to make use of them forgotten:

> If you don't do a method or protocol for one year, then usually it's falling out of
> the daily life /…/ I think this knowledge holds for 1–2 years maximum if it's not
> practiced. (B1)

On the utility of sharing data openly for others to use, one researcher expressed that s/he could only see one purpose of doing it: "It would be of historical interest, to people studying [the project's research topic] invented in 2011" (A). Several researchers offered various practical concerns and obstacles when talking of making data accessible to others. One researcher expressed ambivalence towards providing access to raw data, because even though sharing is important, specific knowledge is needed to use the data:

> It's a fine line between making raw data available to others, which I think is
> important, that people do not only see the evaluated data. *But* it would also
> become very impractical for anybody else to dig through that data, without
> having the knowledge that we have, in order to extract data out of it. (C3)

Other researchers agreed with the standpoint that the openly accessible raw data would not be understood, "it's no use to really open raw data because... [laughs] no one can really interpret them" (B2). To interpret and make raw data useful would require access to the information written in the lab books:

> If we now somehow made our raw data accessible to the public /…/ nobody
> would know what to do with it. This would only make sense *together* with
> basically also publishing our lab books and there again there's something… I
> mean you put personal notes in there, new ideas, stuff like that. (C1)

However, in the DMP's description of data types, source code and device layouts were stated to be of expected utility to researchers in related fields. As for the developed products that the source code would result in, the project stated "We expect this research field to grow even after the end of the project, therefore we expect many researchers to use the softwares [sic!] resulting from this project." Related to this, the researcher working with source code said that s/he had hopes that both tools and the material for developing them would be used and studied by others:

> We would [like] – together with publications – to make some of the algorithms
> available, even the tools that I showed you now. Part of the nice thing about it is
> that when people read the papers they can go to some place and try out to see
> what we did. (F)

One researcher expressed that the project data might be interesting to the few other groups in the world that also use "these particular type of methods" (B1). For the larger scientific community however, the data were considered to be too specialized. One researcher compared the project data with data from particle accelerators. S/he thought the latter could be used to make other interpretations in contrast to this project's data:

> In our case it's much more… I don't want to say sophisticated, but more
> specialized in some way. So I don't know who could really benefit from this,
> from the raw and not interpreted data /…/ Nevertheless they wouldn't really get
> much knowledge from the pure data because not all our personal experience,
> what we really perform in the lab, is documented there, because it's in our
> minds. (B2)

The researchers in the partner group working at a private institute, group B, often carried out contract research with industrial partners, which made them concerned about making their data accessible. They were the only researchers that feared that making data accessible could lead to a violation of the property rights agreements they had signed with their other industrial customers. One of them explained s/he stored the data generated for this project on a server where restricted data generated for these industrial customers also were stored. This procedure made it difficult to separate data belonging to different contracts in a secure way (B2). This researcher's colleagues added, "we cannot open our data files created in industrial projects at the same moment" (B1). Others in this group clarified:

> We are not a university but are involved as a private institute /…/ [that is]
> financing itself from contract research to industry so this is always going under
> some kind of IP directory of property rights management. All that is created in
> project together with customers of course is not allowed to be published /…/ so
> we have to *select* and annotate correctly the data which we can publish.
> (B1)

> It's contract so you're *restricted*. (B3)

> We know that data *created* within [the project] has an obligation to be published
> by the European Union provided the contract we made by stepping into the
> project, so that it has *their* rules. And we know that data created in other projects
> for industrial customers is not allowed by any means to be disclosed. (B1)

Regarding long-term data preservation, several researchers as previously mentioned stored generated data on institutional servers, where they were easily accessible, for periods when they were used continuously. After the data were used for varying purposes, such as a publication, they were moved and stored in a less accessible place:

> Until the project finishes, these microscopy data, the movies, stay on a server
> where they are easily accessible in order to do our data processing on them. And
> afterwards, if we have done all the data processing – usually when we have
> published a paper – they get archived on the same server but on another
> partition which are usually tape drives where it takes longer time to retrieve
> data. (C3)

The same researcher explained that the reason for not discarding the data remaining after extracting the interesting parts, was to guarantee the possibility to control data or go back and discover new things:

> We could also extract the data right away and then throw away the movies,
> 'cause the movies need lots of storage space, but then we would never be able to
> go back. If later on we find out something 'oh maybe there was something
> interesting that we didn't look at the first time' then it's better to be able to go
> back to that. For us this is the best way and the only way we can do it. (C3)

In the data management plan, there was little information about how the project's data were planned to be preserved. The section regarding costs and potential value of long-term preservation reflected how the group saw their data to be of little value for future reuse by people outside the project, as they would quickly become outdated or irrelevant:

> The data generated by the [project name] are probably not useful for third
> parties more than five to ten years after the end of the project due to newer data
> being more relevant. Therefore, we do not plan to ensure data preservation
> beyond that time. (Data Management Plan)

Analytical summary

The material shows that sharing data with people outside the group is viewed as a meaningful activity, seemingly self-evident, within the community: "making raw data available to others /…/ is important" (C3). This understanding is reflected both in the comments of how the participants share data on request, or together with publishing scientific articles ("you always have some supplementary for a paper", C2), or in the published articles to which researchers had added related data. These activities appeared not to be driven by the external policy but came from within the community. A check on the published articles showed that none of the journals where the researchers had published had data policies, but they did however encourage dissemination of data for transparency. It can therefore be assumed that data shared with scientific articles were shared at least relatively voluntarily, as support for the claims made. To share in this way is something they want to do and do independent of external directives, as the computer scientist stated (F).

At the same time, as seen earlier in this chapter, the preoccupations researchers expressed showed understandings that appeared to make them hesitate to make the data openly accessible; to interpret their data was considered to require specific skills to use them: "…without having the knowledge that we have" (C3), or access to the information written down in the lab books (C1), these data cannot be used. Otherwise, both sharing the data and having access to the data become meaningless. Being able to "interpret" data, as B2 pointed out, is once again key for enabling use of the data. In addition to these worries, other statements showed that sharing project data was not seen as meaningful, neither for the community itself nor for others. Data would not be interesting or useful aside from "historical interest" (A), as data are rapidly outdated by technological development (B1 and B3). The fast pace makes perspectives of 5–10 years appear very long for the researchers, who in some cases thought data would be outdated after only one year. When tools change, daily routines change. Activities are quickly forgotten and thereby the knowledge needed to understand the data as well. However, source code data was an exception, and data that some believed would be useful to persons outside of the project. There were expectations that, once made openly accessible, these data would be tested by external persons, which might even lead to communication in the form of comments or feedback. Aside from documentation, several researchers did not see themselves benefitting from re-using their own data. Data appeared to be of little interest to the researchers once they had used that data for publications.

The DMP also showed that the group agreed to make their research data partly accessible and partly closed. The closed data would be accessible only to the group until decided otherwise, and this action was justified because only community members should access these data and use them for the community's purposes. This view is in line with the data policy that allows Intellectual Property Rights as a valid reason for closing data. It was clear that it was not the data type that decided whether data were to be closed, but the purpose for which they could be used. The understanding of certain data as "exploitable" or "protected", that is, legally protected as intellectual property, was key for the community's decision to share them or not. The notion of "exploitable" was explained as contributing to and constituting a part of research publications or, as we saw earlier in the data overview, to patents. To use certain data for these activities and to prioritise protection of intellectual property before sharing them appeared to be very important to this community's members. After having used the data fully for these purposes, the data were less valuable and thus possible to share with others.

The specific concerns of the participants from the private institute illustrate how the researchers in this partner group were also members of another community of practice, in which different procedures and values were accepted and prioritised. Their work situation is on the one hand shaped by the demands of industrial customers with whom they had negotiated agreements regarding data management. On the other hand, their work was also shaped by how the policy was interpreted within this research project, and by how data were agreed to be shared within this community. These researchers stood between competing claims that had to be balanced. They hesitated to share data because of their fear of violating agreements with industry, to whom their commitment appears to be greater.

The co-coordinator's account of how Zenodo had been used in the group's attempt to comply with the data policy and make data openly accessible via a data repository as a response showed that it had not been successful, which, not surprisingly, "discouraged" (C1) the group from trying to use it further. The repository did not suit the group's data; data were too large and did not fit the profile of topics of the repository. The community's attempt to add an explicitly new activity to existing activities ended. Once again, the materiality of the data created made it difficult for the group to share data.

5. Discussion

In this chapter I discuss the empirical material in relation to the theoretical framework and relevant previous studies with the purpose of answering the thesis' research questions. With focus on what is shared and mutually accepted within this particular research group, the group's data practices will be discussed by loosely operationalising Wengers' dimensions, namely mutual engagement, joint enterprise and shared repertoire, which may build a connection between practice and community.

The chapter comprises three sections, the first of which will be about the researchers' *conceptualisations* of data and the first research question. The material for this investigation about researchers' data sharing was collected with the aim of generating empirical material about this subject. During the interviews and observations, however, it became clear that the issue of what data were to the project group researchers and how they conceptualised data in their research work was an unexpected and interesting finding that could not be ignored. One informant, for instance, remarked explicitly that what data was to her/him and the other project researchers probably differed quite a bit from what data meant from my perspective. Examples like this (and there are several more) demonstrated that there was more to discover empirically about this issue than I had first expected. I therefore decided to explore this aspect further, inspired by authors like Cox and Wallis and colleagues (e.g. Cox, 2012a; Wallis et al., 2013), who stated that before one can investigate how people find, use, create and share information, it is necessary to first understand what constitutes information for the social actors of the practice. The assumption is that the premise of Cox and Wallis and colleagues is also valid when studying data sharing in the social practices of researchers.

The discussion about data conceptualisations is to be seen as an important starting point for the subsequent discussion focused on the thesis's main theme, which is researchers' understandings of data sharing. Since there was no research question fomulated on this subject at start and the material was therefore not generated intending to collect information on this subject, the discussion does not in any aspect pretend to be exhaustive, and it will probably raise more questions than answers.

The second section of this discussion will concern the researchers' collective *responses* to the data policy's demands, which addresses the second research question. I focus on the researchers' interpretations of the new requirements, the activities and processes that were initiated in the group when the conditions of the data policy started to apply to project work, and the chosen strategies that guided the researchers. The researchers' definitions of the

circumstances of what has meaning and what does not in relation to data sharing and how they commonly assess appropriate action when shaping responses are at the core, and I aim to elucidate the interconnected understandings at work within the community when the response to the data policy was shaped.

In the third and final section, the answers of the two first research questions will be explicitly related to the researchers' data sharing activities. The third research question focuses on the *implications* of both the researchers' conceptualisations of data and their responses to data policy, and what those ideas imply for how the researchers aim to share data. The aim is to understand the ways in which these notions have consequences for daily research practices, which aspects of data sharing are changed and which are not, or to put it another way, which elements are negotiable or constitutive. I end by taking a step back from the specific results, connecting them to the larger picture and putting them in the context of future perspectives.

After answering the research questions, the chapter will be concluded with critical reflections from various perspectives on the validity and reliability of the investigation's findings.

How are data negotiated and reproduced within the group? What are data to the researchers, and when?

Data are certain and relational, indeterminate and unreflected

The findings showed that there were similarities in both how the participants understood data as concept and in their terminological usage of the word. Some of the researchers' narratives (both experimentalists and theoreticians) described data as being recognised and framed as experimental results and measurements, and by having machines as originators. Data was also characterised by inhabiting qualities as a degree of certainty. This conceptual understanding is evidently firmly rooted in empiricist and positivist traditions and it can thus be viewed as an imported practice, which repertoires often are (Wenger, 1998). This agreement about what data is appears having been formed elsewhere, likely in the training and education that the researchers had received within natural science. This perspective stands in strong contrast to constructionist definitions such as that given by Borgman (Borgman, 2015), because the *use* of data as evidence is not seen as a quality that makes data appear. Rather, from this perspective, data become data in the moment when the machine produces them, which brings them nearer to an understanding of data as objective facts that are given rather than

constructed (cf. Leonelli, 2015). Seeing data as *becoming* data when the machine produces them in addition answers the question of *when* the participants consider data to be. There is thus a temporal factor in this perspective, a moment when data are 'born', a before and an after. In parallel, the strong connection to, or dependency on, machines implies an understanding of data as relational – the machines are what make data be data (cf. Haider & Kjellberg, 2016).

In parallel with this definite and demarcated conceptualisation of data, the findings showed diffused understandings of the concept that are more vague and less easily framed. Data was justified as data by their function, as in the examples with the publications and the lab books and lab journals. These things were seen as data because they constituted shortcuts to other data or gave access to or made other data understandable. Again, this perspective suggests a relational view: data become data because of its/their close connection to more determinate data. It is difficult to fit a publication's function as a shortcut to data into Chao's categories of data as dimension of data or data attribute. In addition, as one member of the research group explained, the practice of including publications as data in the data management plan resulted from the group's will to satisfy the funder's open access policy. However, the understanding of lab books as explaining data could potentially fit as metadata within in the sub-category of "knowledge representation" (Chao et al., 2014). Or, even more suitably, the lab books could be described as *paradata*. A recent study uses this concept to describe data with relation to other data and that – in contrast to provenance data (the datas' origins) and metadata (the datas' kind) – specifies "the past, ongoing and potential processes relating to data" (Huvila, 2022, p. 32).

Furthermore, the term data was used much more flexibly in daily work compared to the above determinate definition. The members accepted and allowed *data* to denote different types of material related to the group's research, all of which were understood as having varying grades of data-ness. The term data was flexible and elastic. This terminology use offers another explanation for why these STEM researchers included certain information, such as metadata and publications, in their wider view of data. This use also suggests a perspective of data that acknowledges a broad range of functions and characteristics that can plausibly fit under the umbrella concept of data. Interestingly, this use aligns with the broad use of the term that has been observed among humanities faculty researchers (Thoegersen, 2018).

Despite the daily use and handling of data, this repertoire was mainly unconscious, implicit, and expressed as unreflected. While the researchers recognised data as data as by a

"scholarly act" (Borgman, 2012, p. 1061), these understandings existed as tacit knowledge. The researchers were not engaged in actions related to enabling negotiation of the data concept or forming common ways of valuating or making sense of data as such. The participants *knew* what data was, as a silent agreement, and this was not a subject that needed discussion or agreement.

Differing data practices coexisted

The results showed that the participants had one of two recognised informal identities, either experimentalists and theoreticians, and that their identites had implications for their understanding of what data was, which also derived from what data was supposed to be. The group is heterogenous and the experimentalists and theoreticians thus formed two bounded communities of practice that coexisted within the larger community, displaying their multimembership. While both groups agreed on the previously discussed determinate view of data, their different identities articulated a disconnect (Wenger, 1998) between them in how terminology was used and how that which they generated was represented materially. The experimentalists used the term data unproblematically and evidently; they undoubtedly worked with data. The theoreticians for their part lacked the words to describe what resources were generated within their community, and the found terminology such as *ideas, designs* or *verifications* more relevant to their activities. To this group, *data* was not a relevant term. This understanding is not too surprising, because they did not think that what they generated was data in the first place, according to the community's shared determinate view; data was measurements, certain, and machine-generated. Understandings of data thereby drew boundaries between the two groups of researchers in the same way that understandings of data have been shown to separate philosophers from other humanities faculty researchers (Thoegersen, 2018). Further, even though it is important to keep in mind that what these two theoreticians generated was not the same, neither of their ideas fit under what is considered a common understanding of data, as numeric and quantitative (Oxford English Dictionary, 2018; Thoegersen, 2018).

The findings also demonstrated differences between the groups of researchers in their material representations of data. If one considers what the mathematician generated *as* data (by applying this thesis' working definition of data), these generated *ideas* completely lacked material representation. In contrast, the material aspects of the data of the experimentalist researchers were repeatedly mentioned throughout the observations (that part of the experimental data was represented by very large computer files). The results nuance existing

knowledge of how data representation practices differ between groups and how these representations are intertwined with other research practices (Borgman, 2015; Chao et al., 2014). The research of experimentalists and theoreticians differs in approach, instrumentation, and community, which helps shape how data are materially represented. Previous research has descibed how the use of data defines boundaries between experimentalists and theoretical modellers (Birnholtz & Bietz, 2003), and has, at least briefly, mentioned variations in how these groups use the term data (Borgman, 2012) without however commenting on differences in the data's material representations.

The community is heterogenous because the community participants were assembled by careful selection based on their competence. All members knew who represented which discipline, and also who belonged to the experimentalist group or the theoretician group. The researchers also knew that being an experimentalist or theoretician meant differences in how data were generated and represented. Belonging to a certain discipline was however something that was mentioned in relation to what data was. Though the material presented examples of situations where this boundary meeting between the different data practices created tensions, the material did not specifically describe how these tensions were overcome. However, researchers' narratives indicated that the co-coordinator's function as a *broker* (Wenger, 1998, p. 109) was highly valued, since this person was able to mediate and link between perspectives at times when members' different practices threatened to weaken the community's coherence.

To summarise, the results have unveiled some of the researchers' perspectives of the *data* included the *data sharing* activities I study; these are their views on the data they share. Even within a relatively small community of practice, an unexpected complexity showed that many interpretations and perspectives on data can exist in parallel. Some of these perspectives were easier to frame and explain while others were more enigmatic. In addition, data practices were found to differ for the two different identities, and this point has not been discussed often. When comparing this complexity to the definition used in the funder's data policy,[18] for example, the contrast is evident. The funder's definition is however relatively inclusive and appears to be able to house the data created both by theoreticians and the experimentalists in this group. However, it does not cover the richness of understandings observed within this community. Thus, as others have pointed out (Borgman, 2015; Haider & Kjellberg, 2016;

---

[18] The European Commission defines data as "information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion or calculation" (European Commission, n.d.-a).

Leonelli, 2015), the concept of data continues to be oversimplified. Data is a dynamic and complex concept, formed by relational, situational, and disciplinary factors as well as by external expectations.

How do researchers mutually account for the data policy?

Sparking learning and engagement

The results showed that a few of the project members, mainly those from management, struggled to understand what had to be done and how it had to be done when initially encountering the data policy at the project's start. They tried to grasp what tasks the data policy would entail for them and how they could fulfil those requirements. Much was unclear and there was little time to dedicate to the work. The encounter clearly highlighted that the community lacked the knowledge necessary in several areas to be able to respond to the policy; this uncertainty thus sparked a learning process. To learn more, researchers depended on help from outside the community. Librarians, peers, and funder support all helped to interpret the policy, bridge between community and policy, and make the requirements understandable. The results showed that researchers learned of technical matters such as metadata standards and interoperable identifiers, but also about suitable data repositories, about potential ways to make data openly accessible, and about how to make project data FAIR in practice. All these issues were evidently new to the researchers.

When project management introduced the data policy and the work that this policy would entail to the rest of the project group, results showed that they consciously chose a motivating approach. Knowing that this approach would entail additional tasks, they emphasised the potential advantages of the data policy work and tried to make the policy meaningful to everyone rather than imposing it on everyone. By striving to present the policy it in a way that felt bearable, they were creating an environment where engagement in forming a response to the policy was made possible (Wenger, 1998).

The findings also showed that, in the process of the concrete work of forming the data management plan's content, all researchers were asked to contribute with their specific knowledge of the data they worked with. In this interdisciplinary group, each participant was considered to best be able to suggest how their particular data potentially could meet the demands for management and sharing. The heterogeneity of the socially complex community

(Wenger, 1998), strategically composed by complementary competence, was utilised, which turned out to be necessary for being able to respond to the data policy, since no single participant alone had all the data-related knowledge that was needed. Every project member was thus involved in the process of shaping a response to the data policy and had to dedicate time and effort for this task.

The group's pursuit of shaping a mutual response to the data policy resulted in the community producing new shared resources. The documents created for overviewing the project data and the data management plan constituted concrete resource objects that manifested the group's specific perspective of dealing with shared data and interpreting the data policy (Wenger, 1998). It is noteworthy that, after the intense first six months when the data management plan was developed and delivered to the funder, there were few subsequent signs of researchers being engaged in activities connected to shaping a response to the data policy. Evidently, the data policy was interpreted as if delivering a data management plan was a main task.

Complying by avoiding change

In the process of shaping a response to the imposed data policy, the group's narratives showed that the group formed an explicit strategy about how this response was to be accomplished: by dedicating as little time and effort as possible to the requirements and by avoiding changing the existing data practices. This result is interesting from many perspectives because the approach clearly demonstrates the researchers' priorities and what they commonly thought was of importance during the process of negotiating a response. The decision to devote as little time and effort as possible to the tasks needed to meet the external requirements was the least surprising issue. The task was evidently involuntary: it was neither initiated by, sought by, nor driven by the participants. As time and effort have previously been found to constitute significant barriers for researchers' decisions to share data (e.g. Borghi & Van Gulick, 2018), it was not unexpected that these factors also had implications for these researchers' approches to the data policy.

More surprising and interesting was the community's explicit strategy to not change the existing data practices, and this observation raises more questions. This way of dealing with the external requirements was considered to be an appropriate approach (Wenger, 1998) by the community. This result stands in contrast to the researcher statements that expressed positive attitudes towards the general idea of making data openly accessible to a wider public. For instance, it was stated that doing so could potentially create new collaborations. More

importantly, this approach contradicted the fact that most researchers engaged in frequent data sharing both within and outside of the community, thereby enacting the activity's importance. Such an existing engagement already in place could raise expectations of more interested and open attitudes. These results can be seen as a gap between idea and practice regarding sharing data openly, and confirm a discrepancy that several other studies have previously found (e.g. Darch & Knox, 2017; Zenk-Möltgen et al., 2018). However, a more plausible interpretation is that the approach to not change the contemporary data sharing practices originated from the researchers' general experiences of data polices being meaningless: a data policy was described as increased bureaucracy taking time from research, and the narratives gave a number of reasonings.

First, managing and sharing data were activities already considered to be highly important and a part of the communication skills a researcher needs to have. To consider in detail aspects of how, when, and with whom to share data were activities that most participants carried out daily as an integrated part of the research practices. These things were obvious and therefore did not have to be pointed out by others. It is worth noting that researchers made little difference between sharing data within and outside of the project; the considerations taken before communicating data with another person were the same. Furthermore, sharing data was seen as an unproblematic activity that was not in need of new solutions. Independently of external requirements, researchers thought data sharing in general worked smoothly within the project, and participants shared data voluntarily with external interested persons on request, along with research publications, or on an online platform for source code data. Nevertheless, they evidently had things to learn about making data openly accessible, as the encounter with the data policy showed. Furthermore, the researchers were preoccupied with the notion that the project data were too specialised for outside people to understand, and that people not involved in the data generation processes could not understand the data. These issues were related both to the fact that the researchers were alone in developing certain research methods and that distance from the research processes was thought to reduce the possibility for understanding the data. To the researchers, it was central that data – if they were to be shared – should be useful to the recipient and thus were delivered well explained. The generated data were also considered uninteresting for others as the group had not yet really solved any problems or made new discoveries; only those 'useful' data were thought valuable for sharing widely and without request. Also, the researchers could not find an available repository that suited the project's generated data. The group therefore had had to settle with a general repository, which was a channel that they doubted would allow their data

to be findable by people who were interested or could benefit, and which, in the end, rejected their data anyhow.

Even though the researchers at an early stage had realised that the funder would allow exploitable data to remain closed until scientific articles had been published or patents had been formed, data sharing remained an issue that apparently concerned many of them, at least when initially encountering the policy. Protecting intellectual property for competitive reasons was clearly highly important to the participants, and was of particular significance to the partner group researchers from the private institute who were used to being bound by contracts with industrial customers. To them, the data policy was seen as a potential threat that could make them violate these agreements involuntarily. This observation shows that these participants belonged to another community of practice in parallel, or in other words, their multimembership (Wenger, 1998). In confronting the demands of these competing practices with different interests, the researchers' commitments to the industrial partner appeared to be stronger than commitments to the research funder; complying with these contracts was more relevant than complying with the data policy connected to this research project. Their loyalty to the industry that financed their home institute appeared make them hesitant to risk industrial contracts merely in order to make data accessible. Another potential explanation is that the consequences of not following the industrial contracts would be more severe than the consequences of not following the research policy.

Several of these results confirm and further nuance previous findings of researchers' attitudes towards data sharing. Sharing data informally with other researchers on request, which was a common way for many of the researchers to share with outsiders, has been shown to be a popular form of sharing (Wallis et al., 2013). That closeness to the data generation processes is highly valued by researchers has previously been identified as an important factor for the ability to interpret data (Borgman, 2015). Further, having data repositories available for sharing data has been found to be a motivator for researchers to share data openly (Federer et al., 2015; Kim & Nah, 2018). In this project, the lack of a suitable repository made the researchers insecure and discouraged. The importance of protecting data before publishing articles or registering patents is a finding that both illustrates the competitive nature of scholarship (Borgman, 2015) and aligns with previous observations to the effect that researchers in several fields prioritise exploiting data fully for their own needs before making them accessible to others (Borghi & Van Gulick, 2018; Fecher et al., 2015; Tenopir et al., 2015). This tendency was found to be particularly strong among

biologists and physical scientists (Tenopir et al., 2015), disciplines to which the project researchers either belonged or were close to as members of STEM disciplines.

All these ways by which participants ascribed meaning to their daily data-related activities and defined what was important and what was not for them, reflect the community's joint enterprise (Wenger, 1998). These practices elucidate the community's approach of not changing existing data practices and explain why participants generally showed such low interest in exploring the data policy's ideas beyond the minimum amount necessary. The result raises the question of why the research group interpreted the data policy as mandatory and why the possibility of opting out of the Open Research Data Pilot was not discussed. One explanation may be that the policy directives were vague and left space for interpretation, and that it was easy to simply make the data policies align with the group's existing practices. Furthermore, the policy was flexible in that it allowed confidential data to be kept closed, in line with the research group's preferences. Because adding a strategy for data when applying for funds was seen to potentially increase the possibility of getting funding, it is reasonable to assume that the group complied with the policy on the assumption that doing so would be favourable for the relationship with the funder generally.

In addition, the material has also shown that dealing with demands from different external parties was routine for most of the participants. The group had had previous experiences of open access policies, and some researchers were used to following industrial contract rules in their daily research work. In this perspective, even though the data policy and its particular demands were new to all members, it appears that, in the end, it was commonly understood as just one among many other directives that simply had to be executed in order to continue with the research. These were all factors that made compliance be carried out in what can be characterised as an instrumental way, a result similar to previous observations of how engineers responded to data policy (Mallasvik & Martins, 2021).

This instrumental approach can be exemplified by the group's approach to the four requirements of the data policy, namely to develop a data management plan, deposit the data in a data repository, ensure third parties can freely access, mine, exploit, reproduce and disseminate the data, and provide related information and identify the tools needed to use the raw data to validate your research (OpenAIRE, 2017). The researchers developed and delivered a DMP after six months, thereby following the funder's rules and fulfilling the first requirement. The group attempted to put produced project data in a research data depository and took necessary action to fulfil this second requirement but was hindered by factors that lay beyond their control (lack of a repository that suited the project's topic and data size). For

these reasons, third parties were not able to freely access, mine, exploit, reproduce and disseminate the community's data, nor was information needed for using the raw data to validate the research provided (requirements 3 and 4) as a consequence of the data policy. Technically, one of the four requirements was thus fulfilled.

In addition, various narratives showed that the data management plan was thought to be helpful by a junior researcher who was new in the research group and in her/his career, since the policy explained how data were to be handled. In reflecting community-specific discourse and explicitly formulating agreed ways of action regarding how the community was to manage and share data, the data management plan functioned as a shared tool helping this newcomer participate in shared practices (Wenger, 1998). Implicit data management routines previously not articulated but tacit (representing the mutual accountability) became accessible and formalised. This implication of data policies is rarely mentioned in the literature.

Finally, the picture of the researchers' policy response as mainly avoiding changing the present practices needs to be complemented with the researchers' expressed intentions and practical attempts to make the project data openly accessible in a data repository, precisely in line with the data policy needs. This attempt to actually change current practice, in contradiction to the group's explicit strategy, showed that the group chose to fully accept external pressures and was willing to explore potential new ways of data sharing.

To summarise, this discussion demonstrates the understandings of data sharing that helped shape this specific research group's priorities during the process of negotiating a mutual response to the requirements of the imposed data policy. The group complied by mostly avoiding changing their existing practices of sharing data, since these were well founded in their needs and the researchers were generally satisfied with them. These researchers furthermore considered themselves to be highly skilled in both data management and data sharing; they knew how to make data understandable to others, which is why they did not think they needed to have such self-evident issues highlighted. The group's response was also formed by their understandings of the project data; the researchers considered the project data to be both uninteresting and potentially incomprehensible to people outside of the group. The researchers' understandings of their work situation (lack of time, loyalties to other parties with competing policies) also shaped their responses to the data policy as did the relative flexibility of the policy directives themselves, allowing room for interpretation and adaptation.

Notwithstanding the researchers' well-developed skills in and experience of data sharing and the fact that they used project members' specific knowledge of the project data, the

encounter with the policy revealed a lack of knowledge necessary for being able to respond to the data policy. External support structures (librarians, funder support) proved to be indispensable for bridging knowledge gaps during the process. However, after the delivery of the data management plan, there was little or no engagement in activities related to the data policy.

How do the views on research data relate to how researchers aim to share data?

To relate the group's understandings of data, the demarcated view of data as measurements and machine-produced or the researchers' flexible use of the term data, to how the researchers aimed to shared data is not easily made and the empirical material was not collected for this scope. I have discussed how the lab books were seen as data because of their function to make experimental data understandable, as meta-data (Chao et al., 2014); these explained the conditions and settings around the experimental processes that had been carried out. To be able to ensure that shared data were made interpretable and meaningful to the recipient was something that the researchers considered to be important. Thus, in this aspect, the lab books formed communications tools which made intellectual exchange possible (Leonelli, 2015). Their contents were instructions from one researcher to another about what had been done and what would be the next step of the experiments and thereby helped enable data sharing within the project.

The fact that the mathematician and computer scientist did not consider that they generated any data, since neither of their ideas fit under what the group thought of as a common understanding of data, did not seem to have implications for how they aimed to share data. There were no signs of that this was a discussed issue. The same is valid for the researchers' flexible use of the term data; it was not a discussed issue because it was unproblematic, or irrelevant, in relation to how they shared data. Again, the researchers *knew* what data was.

Regarding the material aspects of project group's data, these had however significant implications for how the researchers aimed to share data because they conditioned these activities. In the case of the experimentalists' microscopy data, for quite a while, the community members had been having difficulties with finding ways of sharing these data between group members. This difficulty in sharing within the group was for instance one of

the reasons that the researchers were satisfied with the new file sharing solution that was implemented for the current project period: the new system enabled internal sharing of these large data files. The material representation of these experimental data created conditions to which the members adapted their activities, for instance by choosing a suitable channel for sharing or by extracting some of the data from larger files. The datas' physical format also hindered the researchers from making these data openly accessible outside of the community by depositing the data in a data repository, which was the data policy's second requirement. The general repository that the group had chosen initially for this purpose did not provide sufficient server space for uploading the large microscopy data files, and consequently the data were not shared. Thus, both insufficient infrastructure and the material representation of the data formed barriers to the researchers' aim of sharing outside of the project.

The case of the mathematician's data was noteworthy because these data had no material representation. These data were stored in the researcher's mind which had implications for how these were shared. This fact could disrupt both internal research work and the relationships between the researchers in situations where verbal explanations of these ideas did not suffice. The lack of material representation hindered communication and understanding between the participants and meant that the data could not function as a social tool for intellectual exchange (Leonelli, 2015). The mathematician was reluctant to write her/his ideas down because it would limit their content. Thus, rather than changing data by *separating* them from their material form, which is often needed to make data possible to share (Borgman, 2015), their content would in this case be changed by having to *give* them a material form. The mathematician's data were further not mentioned in the data management plan so there were no declared plans for sharing them outside of the project, presumably because it was not possible to declare their file format. This researcher evidently published research results in scientific articles, so it is reasonable to assume that s/he is familiar with describing these ideas in writing and share data in this way. However, following published research results lies outside this study's limitations. The results confirm previous findings that identify theoretical researchers as generators of data that are not consistent in form and difficult to replicate (Borgman, 2012). The results however add another aspect to the view of experimentalist's machine-collected data as easily shared, something that large files are not. To sum up, the physical format of a large part of this community's data is a condition that shapes their portability and, more importantly, their usability, which has implications for their social function within the community and how they are shared.

How do the views on research data policies relate to how researchers aim to share data and how can data sharing be imagined as a constitutive yet negotiable element in interdisciplinary research practices?

In the foregoing discussion, I have presented the researchers' views of the funder's data policy and the many different understandings that helped shape this mutual account: the group regarded the data policy as meaningless. While the research group's encounter with the data policy sparked learning processes and efforts to interpret and shape a mutual response to it, this engagement was guided by the desire to avoid changing the existing data sharing practices, and these efforts ended once the data management plan was delivered. The data policy was viewed as one demand among many others external demands or conditions that the group needed to meet to be able to carry out the research work, and it was adhered to in an instrumental way with the least possible effort. In this perspective, because of these understandings, the results showed that the data policy appeared to have few implications for these particular researchers' aim to share data.

In parallel with these views on the data policy, the group however acknowledged that data sharing was central to this interdisciplinary research community. Data were shared constantly, "all the time", with a high level of awareness and skills. Sharing data was considered an obvious part of the researcher profession and a necessary communication activity. Data were shared within and across disciplinary boundaries, within the project and outside of the project in ways that were meaningful to the group, since doing this was key to the researchers being able to advance the research processes. Technical solutions and mutual agreements had been put in place for making data sharing as highly functioning as possible. These results displayed with full clarity that the data sharing activity constituted an element that was profoundly interwoven in many of the researchers' daily research practices. Data sharing was essential and evident, built in within the group's identity.

While forming a constitutive element interwoven in the researchers' research practices, and independently of the view of the data policy as meaningless, data sharing also showed to be a negotiable activity; the researchers did in fact attempt to make project data accessible in a research data repository. By accepting this new form of data sharing within their present practices, the group thus agreed changing their mutual data sharing practices (Wenger, 1998) according to the imposed data policy's embodied *norm* (Mauthner & Parry, 2013) for data

sharing (i.e., to provide open access to research data). In this aspect, the data policy thereby had implications for the researchers' aim to share data.

The material showed other examples of how data sharing constituted a negotiable element in the research practices. Meeting the external pressures entailed that the group decided to search for information and learn more about issues related to research data sharing, such as new ways of sharing data. To produce the common data management plan, the researchers discussed and then, explicitly and in detail, formulated their mutual agreements about how data were to be shared. The data policy highlighted data sharing to the researchers from a new perspective; even though they had frequently shared data, they had not previously worked with an overview of all the project data or programmed for future common data sharing in this formalised and systematic way. The novelty of the data policy had implications for how the researchers noticed and became more aware of data and data sharing issues, and a small step was thus taken in the direction that the funder stipulated. With this expanded awareness, the researchers were clearly better prepared to share data outside of the group than they had been before encountering the data policy. The new awareness has potential implications for their future decisions.

The results both contrast and confirm previous studies' findings. For the group studied here, the data policy did to a small degree have implications for how the researchers aimed to share data. However, it did not make up an important reason for these researchers' decisions to share data, a result that stands in contrast to the approach to data policy found among researchers in food science and technology (cf. Melero & Navarro-Molina, 2020). The group studied here had an interest in and willingness to share data with people outside of the project regardless of a data policy; the policy was not expected to increase data sharing in contrast to the expectations of a group of astrophysicists (Zuiderwijk & Spiers, 2019). In the group studied here, the presence of a data policy was a much weaker motivating factor for sharing data than several other factors; similar results were found in studies of researchers in neuroscience and biomedicine (Borghi & Van Gulick, 2018; Federer et al., 2015). In this case, the policy affected *how* the researchers aimed to share data (by using a data repository in addition); these results are similar to the way that the funding requirements affected data sharing among engineering researchers (Mallasvik & Martins, 2021).

Limitations and critical reflections

Comment on theoretical framework and representation

The choice of using the community of practice theory as theoretical framework had its strengths and limitations, as became clear when the empirical material was analysed. Compared to general practice theory, community of practice helped draw attention to particular characteristics of the connections formed between group and activities to a higher degree. Using the three dimensions that connect activity and community enabled identification of specific forms of connections. Regarding the researchers' encounter with the data policy and their responses to it, community of practice theory brought out nuances of how the mutual strategies and approaches the researchers chose were negotiated based on how they made data sharing meaningful and how power dynamics worked. The encounter formed a situation in which active production of a local response, or practice, was expected and allowed. This study confirms that this is a circumstance, previously called a problem-solving situation (Cox, 2005), which is well suited for being analysed through a community of practice lens. Viewing the group's conceptualisation of data as part of the group's repertoire allowed investigation of both explicit and unconscious ways of addressing data. These practices however had more the character of routines anchored in researchers' training, not practices that had been actively negotiated, which potentially explains why the theoretical framework offered fewer explanations on this subject. For such situations, the appeal of the idea of reversing the concept into *practices of community* (e.g. Gherardi, 2009) becomes understandable.

When studying a group of people as a community of practice, with the collective in focus, it is not always easy to know the degree to which one succeeds in capturing shared understandings. In other words, it is difficult to ensure that the collective understanding rather than individual understandings is what the results actually represent. Throughout the investigation I have tried to keep the common ground of the researchers at centre and be aware of the closely related representational issues. In the case of this thesis, it complicated matters further that the results to a large extent are based on individuals' sayings, whether expressed in group discussions or individual interviews. Even with a distinct focus on the social nature of knowledge, specific individual statements might potentially have shaped the results disproportionally, which becomes a risk when informants with strong opinions and a

willingness to express them become dominant in the material. However, on the other hand, elaborate statements like those of the mathematician, for instance, contributed depth and details as well as unexpected results.

Simplifying illustrations and complex data activities

In chapter three, the research data life cycle illustration and the motivation for using it to organise my empirical observations was presented. A few methodological reflections regarding this choice need to be made.

That illustration served well to present a detailed display of the many parameters found within the material. The illustration also provided an overview of the data generated from different sources and gave the results a coherent, chronological organisation. In this respect, it was a useful structure for presenting the many variables including contextual factors that were significant for how the relationship between researchers and data is shaped during different phases in this project. However, even if an advantage of the well-known research data life cycle is that it is recognised by a wide public, it became obvious during writing that to separate the various research stages as distinct and to put data in a strictly chronological order was problematic.

To organise the material chronologically entailed that important interview statements of a general character could fit under several of the identified stages. I then had to choose which phase to present them, where they would clarify the most. Another difficulty was that similar statements repeatedly reoccurred and could be related to several of the illustration's phases. An example was the often-stated importance of proximity between data and researcher. These repetitions at times resulted in overlapping discourse. Furthermore, the fact that the illustration is formed by chronological stages made it problematic to fit detailed descriptions researchers gave of their interpretations and understandings. The solution was to place these results where they suited best within the data life cycle, which entailed an unbalanced text in which certain sections had an abundance of material, for instance the first, "Planning for research data", while other sections, like "Data publishing, preservation, and reuse", were less populated with empirical results. This slight unbalance was however mainly a consequence of having studied the researchers mainly during the early phases of their project when little focus was directed on data preservation.

These results provide motivation for questioning the wide use of the research data life cycle, for instance in presentations of data sharing issues for researchers. These are issues that stakeholders like policymakers, service providers, and librarians need to be aware of. The research data life cycle illustration can obstruct themes (beyond the chronological) to be elucidated and conceal the significant complex relationships that surround data sharing practices. Furthermore, data sharing activity between collaborating researchers is rarely mentioned in these illustrations. To most parties involved in the transition to open science, data sharing implicitly means making data available to people who are not involved in the research and who are by definition outside of projects. However, as this thesis has demonstrated, a functioning internal data exchange is a precondition for research work to progress within projects involving several researchers. When the investigated researchers talked of sharing, in many situations they made little distinction between internal and external sharing. In many aspects, but not all, their approach to how internal and external sharing should be done was the same. As a result, both in the data life cycle illustration used here and in most others, data sharing is seen as occurring in a certain stage, that is, *after* data have been generated, processed, and analysed. For this reason, the researchers' sharing of data with people outside of the project had to be placed within the section "Data publishing, preservation, and reuse", which is to say, the last three phases of the cycle.

The description of research and data sharing as linear processes entail they are activities with beginnings and ends. According to this view, data sharing then takes place close to the end of the process, after research has been carried out. Thematic understandings and reoccurring issues, essential for describing the richness of data practices, end up in the background in favour of an instrumental and superficial picture which limits rather than enables understanding data sharing in its real-life context. However, the results have shown that data sharing is more complex and interesting than something that happens only at a project's end. Data sharing is not bound to a specific activity, the final sharing. Researchers start to share data even before the project starts and sharing happens iteratively during the entire research process. Instead, it would be more useful to view both internal and external data sharing as recurrent themes that must be dealt with by researchers at different phases within the research process. Bearing in mind that studies of data sharing focus on sharing of data with external persons, the value of the data sharing that happens within projects needs to be highlighted and discussed. Internal and external sharing should not be set against each other but be seen as a sharing without a beginning and or an end, continuously and constantly ongoing.

Chasing tensions between disciplines

While working with this thesis, the topic slightly altered. Previous studies had found that data practices vary highly among disciplines, which is why I initially expected to find data-related tensions between the collaborating disciplines. These assumptions guided formation of the research questions and design of the study. Focus on the various disciplines represented in the group was strong and formed the main reason for choosing an interdisciplinary group as a case. The aim was to further explore in what ways the disciplinary data sharing practices differed. As mentioned previously, I learned (surprisingly early, in fact very close to my project's start) that this approach had to be abandoned. When meeting the researchers, they completely agreed that this project in every aspect indeed was very interdisciplinary. However, they did not appear to experience this issue as problematic or that it would create tensions for the collaboration.

The clash between the empirical result and the predicted patterns by previous research raised questions as to why were there no tensions or how potential tensions were overcome. The question of tensions created by data sharing between different disciplines is an interesting topic per se, particularly since interdisciplinarity research is widely promoted by funders like the European Commission, but it is one that unfortunately cannot be addressed by this thesis. One possible and simple explanation for why this interdisciplinary collaboration appeared to run so smoothly in general is that the researchers had quite a bit of experience of work in other interdisciplinary contexts. The researchers said that they usually worked with or in close relation to other disciplines in their home laboratories. To apparently all of this group's members, this way of working was more routine than exceptional.

Additionally, from the perspective of a social scientist, the disciplinary boundaries within the STEM areas appear to be flexible. When individual researchers were asked about their discipline, many often gave long answers like "I am by training a material scientist and through biocompatible materials I came to biophysics". Even though the project members undoubtedly represented different disciplines, their knowledge seemed to partly overlap and form a common ground. One researcher expressed her/his view of this interdisciplinary collaboration as "most of them have natural science education anyway so we are roughly trained to think in the same manner /…/ we have a fairly common language" (E). Also, as has already been discussed, the researchers were well aware of both the differences and similarities of the knowledge of their project colleagues. At my observations, the researchers passed the word over to the specialist colleagues, the ones that best knew the discussed issue,

saying "you are the experts here". The roles were clear. That said, my results clearly showed that group still needed consciously developed strategies and at least one person with a brokering role (Wenger, 1998) in order to have a well-functioning communication across the boundaries of disciplines.

The choice of anonymisation

The decision to declare only two of the the disciplines represented in this project group has had implications for this investigation. As previously presented in chapter three, this question was much discussed in the phase of result analysis.

I conclude that the decision to anonymise the disciplines of the majority of the researchers (the experimentalists) in the result presentation was warranted. I had feared that it would be difficult to fully understand the researchers' accounts without the possibility to relate these accounts to previous knowledge of the disciplines. It will of course partly be up to the reader to decide to what degree the accounts are understandable, but to me as an investigator, the decision helped liberate me from preconceptions of disciplines and from focusing on the specific discipline each researcher represented. Therefore the approch enabled a stronger focus on the actual ongoing practices and processes. The exact disciplines also proved to be of less importance than other identities. Even though I was still aware of the discipline of each researcher, reducing the disciplinary focus might have enhanced my sensitivity to other important factors related to data sharing. In addition, because the STEM disciplines have much in common related to how they carry out research, I maintain that anonymisation does facilitate a translation of the analysis into other contingent contexts.

As for the two disciplines that were named in the investigation (mathematics and computer science), each of them were represented by only one researcher, which brings up the issue of representation. Because these two were a minority (theoreticians) in the community, and because their data practices in some aspects proved particularly interesting, their voices might have dominated the discourse. It is important to keep in mind that these individual accounts might not be representative for the entire disciplines they represent. However, these voices enabled indentification and discussion of other attitudes on data and data sharing present within the group.

6. Conclusions

In this final chapter I will sum up the answers to the research questions, take a broader perspective and connect the results to the overall context, and discuss the potential implications for Library and Information Science researchers, practitioners, and other professionals involved in realising the transition to open science. The results disclose new areas to explore, and I will therefore suggest avenues for future research and possible continuations of this study.

This case study focused on the particulars of one research group; even so, it highlighted the complexity of different conceptualisations of data and the diverse uses of the term *data*, even within one small collaborating community (RQ1). The researchers shared a definite conceptualisation of data as experimental results and measurements with machines as originators, at the same time as the term *data* was used flexibly for several types of information. These understandings of what data are were unreflected, shaped by research methods originating from membership in different research identities, by formal trainings, and by assumptions about the expectations of the funding agency. That individual researchers belonged to two different informal but explicit researcher identities, *theoreticians* and *experimentalists,* had implications for how relevant the term *data* was to describe the outcomes of their research work and for how data were represented materially.

This study further provided new in-depth insights, from the researchers' perspectives, into what happens when a research group encounters a data policy for the first time and shape a mutual account for the policy (RQ2). Notwithstanding the researchers' well-developed skills in and experiences of data sharing, they lacked knowledge necessary for being able to respond to the data policy. They therefore initiated a learning process with help from outside the community that could make the policy's requirements understandable. The process of shaping a response to the data policy was characterised by resistance; the group tried to avoid changing their existing data sharing practices. Compliance to policy was thus accepted mainly in order to satisfy the funding agency and be able to continue their research. The data policy was adhered to instrumentally with the least possible effort as one among many other external conditions.

Regarding the researchers' conceptualisations of data, the results showed that their demarcated definition of data and inclusive use of the term data had few implications for how the researchers aimed to share data while the material representations of the data had (RQ3).

The data that lacked material representation hindered data sharing between project researchers and the large files of experimental data could not be shared in data repositories. The data's materiality thus shaped the datas' portability and usability and created conditions to which the members had to adapt their data sharing activities.

When connecting the researchers' views on the data policy to how they aimed to share data, the investigation showed the data policy had few direct implications (RQ3). The data sharing methods indicated by the policy were not anchored within the researchers' practices and understandings of what was meaningful and prioritised, in contrast to the researchers' existing data sharing practices. The group attempted to make project data accessible in a research data repository, but suitable infrastructure was lacking. As a result, only one of the data policy's four requirements was fulfilled: the group delivered a data management plan. The meeting with the external pressures however entailed new discussions and learnings that resulted in an expanded awareness of data and data sharing issues which have potential implications for the researchers' future decisions. The data policy's longer-term implications on data sharing practices should therefore not be disregarded.

Implications

The picture of the unexpectedly multifaceted relationship between researchers and data increases the clarity and granularity in our understanding of how researchers conceptualise and speak of research data. It problematises the use of data as a homogenous and simplified concept that is often used by parties within the knowledge infrastructure, like policymakers. Data is a dynamic concept and what data is to researchers cannot be presumed beforehand. The development of open science and data sharing incentives brings the question "what *is* data?" into the spotlight. It is therefore fundamental that providers of academic support and those who formulate data policies can relate to and be aware of researchers' perspectives on data. Using either inclusive definitions of data or using other broader terms, such as for instance *research material,* would be ways to ensure that data policies are relevant and more reflective of researchers' views. In the future, new categories of data might emerge beyond the current ones, or we might find other concepts and terminologies that reflect and satisfy the needs of the varying data practices of different groups.

Furthermore, the results raise questions of the effectiveness of an imposed data policy (cf. Tenopir et al., 2015) and its potential to change researchers' data sharing activities (Jacob & Hellström, 2018). Policy can, to some extent, help educate and prepare researchers for the transition to the new norm of open science and shrink the gap between an ideal and actual

practice. I however argue that data policies can change little about how researchers share data — particularly researchers who already view data sharing as an essential and meaningful activity, but potentially others as well. Because the meaning of a data sharing practice cannot be separated from the context where it happens, changes in these practices are difficult to alter from the outside. It is true that parties like research funders control the resources on which researchers depend, in a manner similar to how scholars once were bound by the economical arrangements and patronage of kings (David, 2008), and these funding agencies are thereby in the position to challenge renegotiation of practices (Wenger, 2010). Nevertheless, policies cannot directly influence the researchers' practices or *erode* the researchers' control (Mauthner & Parry, 2013) over the decisions about how to share data. If the requirements are not understood by the researchers as meaningful, the data policy will, as has been shown, be complied with instrumentally, in a way that does not entail change of practice. This lack of real change, together with an insufficient infrastructure incapable of providing an appropriate data repository, is the main reason that the data policy in this case failed to realise the ideal of the EU's visions. A potential way to address the tension between researchers' and policymakers' understandings of data sharing and to make the data policy's demands more meaningful could be to encourage researchers to make customised choices for specific data and studies themselves, rather than requiring open access to data.

Even though archiving research data for future needs is not most funders' primary focus, a long-term perspective must be considered as important as speeding up innovation and combating scientific fraud (EUR-Lex: Access to European Union law, 2012) since we cannot know what research data is important to store for future societal challenges. I argue that policymakers need to better communicate the idea of data's potential and unimagined future value, which may have the potential to motivate researchers to put the existing positive approach to the *idea* of open data into practice.

Reforming the research system and how researchers and research are assessed could also motivate researchers to share data when sharing is not research driven, and thereby make these efforts worthwhile and meaningful. Participation in the Open Research Data Pilot did not reward researchers: producing patents or publishing research results does. In the competitive environment of research, some sort of formal acknowledgement of data sharing could motivate researchers to share data. It should however be noted that, even in a system that rewards data sharing, it is improbable that this system alone would make what many (peer researchers, industry, and commercial players) consider to be the most valuable data be made accessible in the near future. As long as policymakers allow researchers to prioritise

145

protecting data for competitive reasons, secrecy will still be a persistant force at work, just as it was in ancient Greece and medieval Italy (David, 2008), and society as a whole will not benefit from these data as was intended.

Recommendations for further research

Since this investigation had a specific scope, the empirical material was not sufficient to thoroughly explain in detail all that data can be to researchers, but it does suggest a number of new pathways for subsequent researchers to follow. Further studies that directly focus on researchers' views of data are needed to expand our knowledge about the relationship between data and researchers. The material aspects of the data of various research groups and the implications of these for how data can be shared is a highly interesting but unexplored research path.

Because the duration of a doctoral project is limited, the data collection for this thesis was limited mainly to the group's second project year (of five and a half years), and the material does not account for how the researchers used the data management plan, shared data, or understood the data policy during the final years of the project. These and the long- and short-term implications of data policies are issues that remain for future investigations in Library and Information Science where data practices is still a novel and emerging research area. Furthermore, similar in-depth case studies of research communities within other areas, for instance social sciences, or of larger research projects or of individual researchers are required so that we can compare different forms of data sharing practices and negotiation of data policy.

Contributions

This investigation has demonstrated the appropriateness of a case study approach to studies of complex, real-life phenomena. With inspiration from ethnography, and by integrating diverse data sources (including interviews with nearly all members, on site observations, and locally produced documents), it became possible to create a rich picture of the coexisting data-related patterns and complexities of one specific real-world system. Interpreting the particulars of this *one* system, as embedded in the larger context in which it developed, paved the way for developing a deeper knowledge of how understandings and responses are negotiated within the group, and how these responses and understandings are shaped by the conditions of the surrounding world. While the generalisation of specific understandings obtained by studying a

single case is limited, the results nevertheless improve our understanding of research data sharing.

As a theoretical contribution, the study successfully developed the Community of Practice concept by operationalising it to analyse the activities of an interdisciplinary research collaboration. The theoretical framework was proven useful for highlighting the shaping of mutual understandings of data – not as homogenic, but as interconnected – and of data sharing within a specific setting.

This investigation has shown the complexity that awaits the project of realising the vision of open data and has provided an example of how making data openly accessible involves understandings of data, researchers' social practices, knowledge infrastructures, competition and collaboration, data policy, economics, and power. This reality is what researchers confront as they navigate among priorities and competing claims. The difficulty of encapsulating the researchers' views of data has demonstrated the importance of investigating researchers' data conceptualisations when studying how researchers share data, and this contribution is clearly relevant for Library and Information Science researchers. This thesis' nuanced picture of researchers' data conceptualisations and data sharing provide policymakers, research funders, developers and providers of academic support with valuable insights that help them fulfil their roles as bridges between data policy and researchers in the transition to open science. Additionally, knowing more about interconnected data sharing practices deepens our understanding of collaborative research.

Epilogue

On becoming a data collector

In relation to my professional role as a librarian, I would like to conclude the thesis with some reflections on how my own attitude towards making my data openly accessible evolved during thesis work and continues to do so. From the start, I was determined to make all collected material openly accessible once the project was finished, probably in a data repository that suited the subject. This noble idea of sharing results and data openly was appealing, and I would gladly contribute with data if they could be of use to others. However, I found that this attitude stemmed from my previous professional role.

After having collected my own data, things changed as I experienced myself much of what literature discusses about the relationship between researcher and produced data, and things that I had heard other researchers say, including those interviewed. Things got more complex and nuanced and even emotional. I felt a heavy responsibility for the collected data that included both the aspect of assuring the data were safely stored for my research purposes, as well as keeping the promised confidentiality to the informants. The data were (and still are) very valuable to me at the same time that they were vulnerable. The thought of losing an audio recorded interview was terrifying and would be an irreplaceable loss that would damage and delay the research process. The feedback on the methods for material collection from the Norwegian Centre for Research Data where the project was registered underlined that the material I was about to collect would include personal data. These materials were therefore, even if not technically sensitive, regulated by law. It was also pointed out to me that sound recordings had to be either completely anonymised or erased after the end of the project.

Above all, I think it was the new experience of carrying out the interviews that gave me a new perspective and made me more hesitant about making data openly accessible, for two reasons. Firstly, the interview is a situation built on trust where the informants give frank accounts contingent upon the promise of confidentiality. It is not so much a question of revealing their identity, since transcripts would be anonymized, but rather one of an emotional kind: the persons shared their thoughts generously with me after a promise had been made. This observation brings up other methodological questions. Would the informants have accepted participation had I made it clear that the interviews would be openly accessible afterwards? And would these interviews have been the same? Would *I* have acted in the same way, as an interviewer?

There are solutions to these problems, such as sharing the material with researchers after the interviews to let them approve and decide what can be made openly accessible and what should not. Considering this possibility in relation to these particular project researchers within this study, this arrangement would have been very difficult to realise. It is likely that they would not have had the time nor interest in participating in reviewing the transcripts. It is also hard to imagine the interviews evolving organically as they did, had it been declared they would be made openly accessible at a later stage. I think these factors might have had an inhibiting effect, both on informant and interviewer. However, all practices and norms are constantly in movement.

The second reason for changing to a less decisive attitude towards sharing the data openly emerged during analysis of the transcribed interviews. I asked myself how these data could be of interest or use to anyone else, beyond reasons of validation and transparency. They are resources generated in line with many, many parameters of choice; the research questions, the methods used, and the theoretical angle all form a very narrow and specifically designed data collection. How would anyone be able to understand them, let alone to reuse them for another study without having been there to produce them? And how could they ever fit to be merged together with other data?

Having now finished the doctoral studies, I have however reached a conclusion. Since the future is unknown to all of us, I can neither imagine what will be interesting for future researchers nor what purposes the data I have collected could serve. Maybe a future historian would like to study research collaboration forms 20 years from now, or perhaps a linguist take interest in comparing interviewing or transcription techniques a century from now. We cannot know of the value of data long after they have been collected or lost (Borgman, 2015). So, I might make the effort and try to put the interview transcriptions somewhere openly for others to use. When I have time.

References

Ahrne, G., & Svensson, P. (2015). *Handbok i kvalitativa metoder*. Stockholm: Liber.

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation, 8*(2), 5-26. doi:10.2218/ijdc.v8i2.263

Alvesson, M., & Sköldberg, K. (2017). *Tolkning och reflektion : vetenskapsfilosofi och kvalitativ metod*. Lund: Studentlitteratur.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., . . . Wouters, P. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal, 3*, 135-152. doi:10.2481/dsj.3.135

Birnholtz, J. P., & Bietz, M. J. (2003). *Data at work: supporting sharing in science and engineering*. Paper presented at the GROUP'03, Sanibel Island, FL, United States.

Björk, B.-C. (2007). A model of scientific communication as a global distributed information system. *Information Research, 12*(2), 1-48.

Borghi, J. A., & Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLOS ONE, 13*(7), e0200562-e0200562. doi:10.1371/journal.pone.0200562

Borgman, C. L. (2007). *Scholarship in the digital age : information, infrastructure, and the Internet*. Cambridge, Mass.; London: MIT.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. doi:10.1002/asi.22634

Borgman, C. L. (2015). *Big data, little data, no data : scholarship in the networked world*. Cambridge, Massachusetts: The MIT Press.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries, 16*(3), 207-227. doi:10.1007/s00799-015-0157-z

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351-360. doi:10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Byström, K., & Pharo, N. (2019). Information artefacts. In K. Byström, J. Heinström, & I. Ruthven (Eds.), *Information at work : information management in the workplace*. London: Facet Publishing.

Carroll, A. J., & Mallon, M. N. (2021). Using digital environments to design inclusive and sustainable communities of practice in academic libraries. *Journal of Academic Librarianship, 47*(5), 1-10. doi:10.1016/j.acalib.2021.102380

CERN Data Centre & Invenio. (n.d.). About Zenodo. Retrieved 8 September 2021 from https://about.zenodo.org/

Chao, T. C., Cragin, M. H., & Palmer, C., L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and

curatorial processes. *Journal of the Association for Information Science and Technology, 66*(3), 616-633. doi:10.1002/asi.23184

Chawinga, W. D., & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research, 41*(2), 109-122. doi:10.1016/j.lisr.2019.04.004

Cox, A. M. (2005). What are communities of practice? A comparative review of four seminal works. *Journal of Information Science, 31*(6), 527-540. doi:10.1177/0165551505057016

Cox, A. M. (2012a). An exploration of the practice approach and its place in information science. *Journal of Information Science, 38*(2), 176-188. doi:10.1177/0165551511435881

Cox, A. M. (2012b). Information in social practice: A practice approach to understanding information activities in personal photography. *Journal of Information Science, 39*(1), 61-72. doi:10.1177/0165551512469767

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology, 68*(9), 2182-2200. doi:10.1002/asi.23781

Darch, P. T., & Knox, E. J. M. (2017). Ethical perspectives on data and software sharing in the sciences: A research agenda. *Library & Information Science Research, 39*(4), 295-302. doi:10.1016/j.lisr.2017.11.008

David, P., A. (2008). The Historical Origins of 'Open Science': An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution. *Capitalism and Society, 3*(2), 1-103. doi:10.2202/1932-0213.1040

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of qualitative inquiry*. Los Angeles: Sage Publications.

Dourish, P. (2017). *The stuff of bits : an essay on the materialities of information*. Cambridge, Massachusetts: The MIT Press.

Edwards, P. N. (2010). *A vast machine : computer models, climate data, and the politics of global warming*. Cambridge, Mass.: MIT Press.

EUR-Lex: Access to European Union law. (2012, 27 March 2021). Commission Recommendation of 17 July 2012 on access to and preservation of scientific information. *Official Journal of the European Union.* Retrieved 22 January 2022 from https://eur-lex.europa.eu/legal-content/SV/TXT/?uri=OJ:L:2012:194:TOC

European Commission. (2016). Open Innovation, Open Science, Open to the World: a vision for Europe. Retrieved 8 August 2019 from https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world-vision-europe

European Commission. (2018). HORIZON 2020 IN FULL SWING — Three Years On - Key facts and figures 2014-2016. Retrieved 8 August 2019 from https://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/h2020_threeyearson_a4_horizontal_2018_web.pdf

European Commission. (2020a). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data. Retrieved 12 March 2020

from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

European Commission. (2020b). H2020 Programme Guidance document. Support for the Research and Innovation Dimension of European Universities. Version 1.0, 26 March 2020. Retrieved 29 November 2021 from https://ec.europa.eu/research/participants/data/ref/h2020/other/guides_for_applicants/other-doc_h2020-iba-swafs-support-1-2020_h2020-iba-swafs-support-2-2020_en.pdf

European Commission. (n.d.-a). Participant Portal H2020 Online Manual: Data Management. Retrieved 21 January 2019 from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

European Commission. (n.d.-b). Participant Portal H2020 Online Manual: Open access & Data management. Retrieved 19 May 2021 from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

Fecher, B., Friesike, S., Hebing, M., Linek, S. B., & Sauermann, A. (2015). A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing. *Entrepreneurship & Economics eJournal*.

Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PLOS ONE, 10*(6), e0129506. doi:10.1371/journal.pone.0129506

Fejes, A., & Thornberg, R. (2015). *Handbok i kvalitativ analys*. Stockholm: Liber.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219-245. doi:10.1177/1077800405284363

FOSTER consortium. (n.d.). Open Science Definition. Retrieved 17 January 2022 from https://www.fosteropenscience.eu/taxonomy/term/100

Frodeman, R., Klein, J. T., & Pacheco, R. C. S. (2017). *The Oxford handbook of interdisciplinarity*. doi:10.1093/oxfordhb/9780198733522.001.0001

GDPR. (2016). Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. *Official Journal of the European Union*. Retrieved 13 January 2021 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Gherardi, S. (2009). Community of Practice or Practices of a Community? In S. Armstrong & C. Fukami (Eds.), *The SAGE Handbook of Management Learning, Education and Development* (pp. 514-530). SAGE Publications.: SAGE Publications.

Gullbekk, E., & Byström, K. (2019). Becoming a scholar by publication – PhD students citing in interdisciplinary argumentation. *Journal of Documentation, 75*(2), 247-269. doi:10.1108/JD-06-2018-0101

Haider, J. (2011). The environment on holidays or how a recycling bin informs us on the environment. *Journal of Documentation, 67*(5), 823-839. doi:10.1108/00220411111164691

Haider, J., & Kjellberg, S. (2016). Data in the making. In J. V. Rekers & K. Sandell (Eds.), *New big science in focus : perspectives on ESS and MAX IV*. Lund: Lunds universitet.

Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm : data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

Concordat on Open Research Data, (2016).

Hills, P. J. (1983). The scholarly communication process. *Annual Review of Information Science and Technology, 18*, 99-125.

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, Calif.: Sage.

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*(1), 28-48. doi:doi:10.1515/opis-2022-0129

Isah, E. E., & Byström, K. (2020). The mediating role of documents: information sharing through medical records in healthcare. *Journal of Documentation, 76*(6), 1171-1191. doi:10.1108/JD-11-2019-0227

Jacob, M., & Hellström, T. (2018). Epistemic governance and the conditions for knowledge production in HER institutions. *Studies in Higher Education, 43*(10), 1711-1717. doi:10.1080/03075079.2018.1520413

Johannessen, A., Tufte, P. A., Christoffersen, L., & Nilsson, B. (2020). *Introduktion till samhällsvetenskaplig metod*. Stockholm: Liber.

Ju, B., & Kim, Y. (2019). The formation of research ethics for data sharing by biological scientists: an empirical analysis. *Aslib Journal of Information Management, 71*(5), 583-600. doi:10.1108/AJIM-12-2018-0296

Jubb, M. (2013). Introduction: Scholarly communications - disruptions in a complex ecology. In M. Judd & D. Shorley (Eds.), *The Future of Scholarly Communication* (pp. xiii-xxxvi). London: Facet Publishing.

Jubb, M., & Shorley, D. (2013). *The Future of Scholarly Communication*. London: Facet Publishing.

Justesen, L., & Mik-Meyer, N. (2012). *Qualitative research methods in organisation studies*. København: Hans Reitzels Forlag.

Kim, Y., & Burns, C. S. (2015). Norms of data sharing in biological sciences: The roles of metadata, data repository, and journal and funding requirements. *Journal of Information Science, 42*(2), 230-245. doi:10.1177/0165551515592098

Kim, Y., & Nah, S. (2018). Internet researchers' data sharing behaviors: An integration of data reuse experience, attitudinal beliefs, social norms, and resource factors. *Online Information Review, 42*(1), 124-142. doi:10.1108/OIR-10-2016-0313

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research, 37*(3), 189-200. doi:10.1016/j.lisr.2015.04.006

Kvale, L., & Pharo, N. (2021). Understanding the Data Management Plan as a Boundary Object through a Multi-stakeholder perspective. *International Journal of Digital Curation, 16*(1), 16. doi:10.2218/ijdc.v15i1.729

Latour, B., & Woolgar, S. (1986). *Laboratory life : the construction of scientific facts*. Princeton, N.J.: Princeton Univ. Press.

Lave, J., & Wenger, E. (1991). *Situated learning : legitimate peripheral participation*. Cambridge: Cambridge Univ. Press.

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science, 82*(5), 810-821.

Linek, S. B., Fecher, B., Friesike, S., & Hebing, M. (2017). Data sharing as social dilemma: influence of the researcher's personality. *PLOS ONE, 12*(8), e0183216. doi:10.1371/journal.pone.0183216

Macnaghten, P., & Myers, G. (2007). Focus groups. In C. Seale, G. Gobo, J. F. Gubrium, & a. D. Silverman (Eds.), *Qualitative research practice* (pp. 65-79). London: Sage.

Mallasvik, M. L., & Martins, J. T. (2021). Research data sharing behaviour of engineering researchers in Norway and the UK: uncovering the double face of Janus. *Journal of Documentation, 77*(2), 576-593. doi:10.1108/JD-08-2020-0135

Mansour, A. (2020). Shared information practices on Facebook. *Journal of Documentation, 76*(3), 625-646. doi:10.1108/JD-10-2018-0160

Mauthner, N. S., & Parry, O. (2013). Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology, 27*(1), 47-67. doi:10.1080/02691728.2012.760663

Meadows, A. J. (1974). *Communication in science*. London: Butterworth.

Meadows, A. J. (1998). *Communicating research*. San Diego: Academic.

Melero, R., & Navarro-Molina, C. (2020). Researchers' attitudes and perceptions towards data sharing and data reuse in the field of food science and technology. *Learned Publishing, 33*(2), 163-179. doi:10.1002/leap.1287

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative Research : A Guide to Design and Implementation* (4th ed.). Newark, United States: John Wiley & Sons.

Morgan, S. L. (2011). Social Learning among Organic Farmers and the Application of the Communities of Practice Framework. *The Journal of Agricultural Education and Extension, 17*(1), 99-112. doi:10.1080/1389224X.2011.536362

Nicolini, D., Gherardi, S., & Yanow, D. (2003). *Knowing in organizations : a practice-based approach*. Armonk, N.Y.: M.E. Sharpe.

O'Reilly, K. (2009). *Key concepts in ethnography*: SAGE.

OpenAIRE. (2017). What is the Open Research Data Pilot? Retrieved 11 June 2018 from https://www.openaire.eu/what-is-the-open-research-data-pilot

Oxford English Dictionary. (2018). "data, n.". *Oxford English Dictionary*. Retrieved 5 February 2022 from http://www.oed.com/view/Entry/296948?rskey=LQHNVr&result=1&isAdvanced=false

Padgett, D. (2017). *Qualitative methods in social work research*. Los Angeles, Calif.: Sage.

Paisley, W. J. (1968). Information needs and uses. *Annual Review of Information Science and Technology, 3*, 1-30.

Palmer, C. L., & Cragin, M. H. (2008). Scholarship and disciplinary practices. *Annual Review of Information Science and Technology, 42*, 165-212.

Pickard, A. J. (2017). *Research Methods in Information* (2 ed.). London: Facet Publishing.

Pilerot, O., Hammarfelt, B., & Moring, C. (2017). The many faces of practice theory in library and information studies *Information Research, 22*(1).

Pilerot, O., & Lindberg, J. (2018). Inside the library: Academic librarians' knowing in practice. *Journal of Librarianship and Information Science, 50*(3), 254-263. doi:10.1177/0961000618769970

Piwowar, H. A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLOS ONE, 6*(7), e18657. doi:10.1371/journal.pone.0018657

Prosser, D. C. (2013). Researchers and scholarly communications: an evolving interdependency. In D. Shorley & M. Jubb (Eds.), *The future of scholarly communication* (pp. 39-50). London: Facet Publishing.

Redden, J., Dencik, L., & Warne, H. (2020). Datafied child welfare services: unpacking politics, economics and power. *Policy Studies*, 1-20. doi:10.1080/01442872.2020.1724928

Schatzki, T. R. (2002). *The site of the social : a philosophical account of the constitution of social life and change*. University Park, Pa.: Pennsylvania State University Press.

Silverman, D. (2017). *Doing qualitative research*. Los Angeles: SAGE.

Smith, S. U., Hayes, S., & Shea, P. (2017). A critical review of the use of Wenger's Community of Practice (CoP) theoretical framework in online and blended learning research, 2000-2014. *Online Learning, 21*(1), 209-237. doi:10.24059/olj.v21i1.963

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, Calif.: Sage.

Talja, S., & Hartel, J. (2007). Revisiting the user-centred turn in information science research: An intellectual history perspective. *Information Research, 12*.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE, 6*(6). doi:10.1371/journal.pone.0021101

Tenopir, C., Christian, L., Allard, S., & Borycz, J. (2018). Research Data Sharing: Practices and Attitudes of Geophysicists. *Earth and Space Science, 5*(12), 891-902. doi:10.1029/2018EA000461

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., . . . Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE, 10*(8), e0134826. doi:10.1371/journal.pone.0134826

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., . . . Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE, 15*(3), e0229003. doi:10.1371/journal.pone.0229003

Thoegersen, J. L. (2018). "Yeah, I Guess That's Data": Data Practices and Conceptions among Humanities Faculty. *Libraries and the Academy 18*(3), 491-504. doi:10.1353/pla.2018.0030.

Thoegersen, J. L., & Borlund, P. (2021). Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. *Journal of Documentation, 78*(7), 1-17. doi:10.1108/JD-01-2021-0015

Tracy, S. J. (2019). *Qualitative Research Methods : Collecting Evidence, Crafting Analysis, Communicating Impact*. Newark, United States: John Wiley & Sons, Incorporated.

Van Tuyl, S., & Whitmire, A. L. (2016). Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia. *PLOS ONE, 11*(2), e0147942. doi:10.1371/journal.pone.0147942

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE, 8*(7), e67332. doi:10.1371/journal.pone.0067332

Wenger-Trayner, E., & Wenger-Trayner, B. (2014). Learning in landscapes of practice : a framework. In M. Fenton-O'Creevy, S. Hutchinson, C. Kubiak, & B. Wenger-Trayner (Eds.), *Learning in landscapes of practice : boundaries, identity, and knowledgeability in practice-based learning* (pp. 13-29). Abingdon: Routledge.

Wenger-Trayner, E., & Wenger-Trayner, B. (2015). Introduction to communities of practice : a brief overview of the concept and its uses. Retrieved 16 July 2021 from http://wenger-trayner.com/introduction-to-communities-of-practice/

Wenger, E. (1998). *Communities of practice : learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wenger, E. (2004). Knowledge management as a doughnut. *Ivey Business Journal*(Jan/Feb). Retrieved from https://iveybusinessjournal.com/publication/knowledge-management-as-a-doughnut/

Wenger, E. (2010). Communities of practice and social learning systems: the career of a concept. In C. Blackmore (Ed.), *Social learning systems and communities of practice* (pp. 179-198). London: Springer.

Wenger, E., Trayner, B., & de Laat, M. (2011). *Promoting and assessing value creation in communities and networks: a conceptual framework* (Vol. 18). The Netherlands: Ruud de Moor Centrum.

Womack, R. P. (2015). Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. *PLOS ONE, 10*(12), e0143460. doi:10.1371/journal.pone.0143460

Yin, R. K. (2018). *Case study research and applications : design and methods*. Thousand Oaks, California: SAGE.

Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of Documentation, 74*(5), 1053-1073. doi:10.1108/JD-09-2017-0126

Zenk-Möltgen, W., & Lepthien, G. (2014). Data sharing in sociology journals. *Online Information Review, 38*(6), 709-722. doi:10.1108/OIR-05-2014-0119

Zimmerman, A. S. (2008). New Knowledge from Old Data. *Science, Technology, & Human Values, 33*(5), 631-652. doi:10.1177/0162243907306704

Zuiderwijk, A., & Spiers, H. (2019). Sharing and re-using open data: A case study of motivations in astrophysics. *International Journal of Information Management, 49*, 228-241. doi:10.1016/j.ijinfomgt.2019.05.024

Appendix I

The interview guide

My focus lies within the area of Scholarly communication, more precisely on research data and researchers' data sharing practices, that is "how you work with data and share them".
I interview you as being members of [project name] but please feel free to answer from all of your experiences from carrying out research.
The interview will be audio recorded. The interviews will be anonymized and only accessed by me and my supervisors. However, the project co-ordinator has approved for me to present the project name which means that this may occur in the final thesis.

- To begin, how did you get involved in [project name]?

- What discipline or disciplines would you ascribe yourselves to?

- I'm curious of your data. Did you bring some data today? Explain what we see! / Can you give me some examples of your data- of what does it consist? If I was to look at your data what would I see?

- How was this data produced? By which method and by whom?

- Where do these data live? How was this done – was some form of procedure needed to do that (was it changed for storing, annotations or settings added)?

- Why are the data put in this place?

- What do you think of this solution?

- If you think of the last week or so, have you shared data with another researcher? Can you describe this situation to me, step by step? Within or outside the project?

- The [project name] is an international and an interdisciplinary research project. Listening to your discussions, you seem to mostly communicate easily even if there were times when you did not understand each other across disciplines. How is it to collaborate around data in this heterogeneous environment?

- Have you shared or do you expect to share data with project members from other disciplines? Tell me about it! How was that? What are your expectations on how that will work?

- If we go back to [project name] and the data produced by the project. Do you think all project participants understand the data of this group (the vocabulary, methods, and theories)? Did you understand all of the presentations in [conference town]?

If we now focus on questions regarding the fact that the [project name], as a Horizon 2020 project, is a part of the EU Open Research Data Pilot. This means the project has had to develop a Data management plan and later make data openly accessible. I have read this plan.

- Have you had previous experiences from research where the funders demanded data management plans? Have you encountered data policies before?

- Did you participate when the data management was formulated last spring? What do you remember? Can you describe the process for developing the plan?

- If I ask you to think of your personal daily research work, in your experience – does the data management plan have implications or effects on your work? Does it make a difference for how you collect, name, share, save data?

- Do you think the data management plan has effects for the collaboration between researchers in some way?

- Compared with how you work with data in other projects, do you handle data differently working with the [project name] data?

- What are your thoughts and feelings on having to develop a data management plan or having to work according to a data plan? Does it have importance? Advantages or disadvantages?

- We have come to the end of the interview, is there anything you would like to add?

I am very satisfied with the results.
Thank you for contributing to my research.

Appendix II

The survey

Have you shared [project name] data?

Your name:

_____

Have you until now made [project name] data openly accessible, formally or informally?

(For example: uploaded on GitHub or Zenodo, added as supplemental information to a journal publication, emailed researcher outside [project name], published in a data journal or put in an institutional repository)

☐   Yes

☐   No

If yes, please specify where below:

_____

_____

_____

Thank you for contributing to further knowledge on scholarly communication!

Regards,

Madeleine Dutoit, Oslo Metropolitan University

A list of errata from the Committee that should be considered:

P.19, bottom: 'written' or 'published' after 2014?

P. 22 'said' should be 'wrote'

Note also that the reference to the study (Thoeghersen, 2018) is mentioned at least trice (P.21, p.26, p.31), in the introduction, without being related to each other. This should be harmonized in the text since this is almost the only earlier study where the Humanities is mentioned).

P.27 "quite as broad" remove 'as'

P. 25 'This' is it 'Chao's study' that is mentioned?

p. 56 I suggest that "*As* not all partner groups…" should be "*Since* not all…"
-"- "…consists *by*" changed to "consists *of*".

There is some recurring information repeated without acknowledging that it has been given before (e.g., the point about Humanities/(Thoeghersen, 2018) above),

This also applies to the mention that 37 (p.30) and 37 % (P.33) of the data sets could be accessed in the study. Which is correct, the number or the share?

The term community of practice poses problems when stated in plural. Different 'community of practices' or 'communit-*ies* of practice' (p.44)? A bit further down, it is stated as 'communi*ties* of practice'. This is inconsistent.

At least one author is not adequately introduced, i.e. Yin on pg 47. While not all authors need to be introduced, those that play a significant role do. For instance, it is crucial if this refers to a textbook author or if the argument stated derives from the methods section in another empirical study.