Enhancing Open Research Data Sharing and Reuse via Infrastructural and Institutional Instruments: a Case Study in Epidemiology

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Complex Systems Engineering and Management

Faculty of Technology, Policy and Management

by

Berkay Onur Türk

Student number: 5221021

To be defended in public on July 5th 2022

Graduation committee

Chairperson and Second Supervisor: Prof.dr. F.M. Brazier, Systems Engineering
First Supervisor: Dr. A.M.G. Zuiderwijk- van Eijk, Information and Communication Technology

Preface

My educational adventure at TU Delft ends with the completion of this master thesis project. When I started in September 2020, we were in the middle of a pandemic with no end in near sight. As I attended my first lecture in front of a screen, I was also trying to stop worrying about the uncertainties of the future ahead of me. Over time, these concerns were replaced by different, more positive emotions. I met many incredibly talented, self-confident, and passionate people who all inspired me to understand what my values are, what I want to accomplish, and most importantly, that I have the potential to achieve my dreams.

I would never have been able to study at TU Delft because of my lack of financial resources. I would like to express my gratitude to TU Delft for making my dream come true by generously offering a scholarship financed by the legacy of Justus and Louise van Effen, who envisioned that technological advancements can greatly address societal issues. Now as I leave this school, I promise to carry this vision for the rest of my life.

As a person who is passionate about social equality, I am especially proud to have worked on a thesis that essentially searches for ways to make data accessible to everyone regardless of where they are in the world. I believe open data sharing can remove barriers in front of many researchers who may not have the same resources as others. I was so lucky that I did not feel alone, helpless, or hopeless in any part of this project because I had a supervisor like Anneke who gave me feedback whenever I needed help, who always allowed me to improve myself, and most importantly, who prioritized my wellbeing during all this process. Anneke, it was a privilege to work under your supervision on this project. I had reservations when starting this project as I had not done qualitative research before, but all of these concerns disappeared thanks to your remarkable guidance. Frances, thank you so much for the guidance you gave to strengthen my thesis, I believe your supervision enabled me to challenge myself in this project.

Finally, I would like to thank my amazing family and friends for their support for the last two years of my life. Mom and Dad, thank you for always being my rock and for teaching me how to act with integrity. After receiving this diploma, as an engineer, I will continue to preserve the moral principles you taught me. Ruben, thank you so much for being who you are, for being my greatest support, and for always standing next to me even on my worst days. You have made me want to believe in myself and to be a better person. Aslı, thank you for teaching me what loyalty looks like in a friendship. Selin, thank you for picking up the phone in another part of the world, almost every day, to remind me how far I have come in my life and that I should always be proud of it.

Please enjoy reading this report.

Berkay Onur Türk
June, 2022

Executive Summary

The amount of data that is collected, analyzed, and stored by researchers has increased drastically in the last decades due to improvements in communication technologies. The 21st century has led to the data-intensive scientific discovery paradigm, which foresees that scientific advancements will be facilitated by increased sharing of data and collaboration among scientists. In line with this, open data practices gain the utmost value for research. Open research data sharing refers to publishing research data on the internet in a freely accessible, usable, modifiable, and sharable format to other researchers. Open research data sharing has many benefits to researchers and scientific fields. It can bring more transparency to the research, save researchers' time by preventing repetitive data collection processes and lead to more collaborations. Recognizing these benefits, across the world, governments, funding agencies, universities, and journals try to increase data sharing practices by introducing new policies. However, despite these attempts, (open) data sharing has not become a prevalent practice in all research fields. This could be due to many inhibiting factors, ranging from insufficient data management support to a lack of data standards for making data interoperable. Acknowledging that the issues in front of open research data adoption differ by field, conducting discipline-specific studies is necessary to reach higher open research data practices in these fields. One field that has relatively lower levels of (open) data sharing practices is Epidemiology. Epidemiology examines health-related phenomena in populations, and it is concerned about finding the causes of such occurrences. Despite the massive amount of Covid-19 related research published during the pandemic, very few of these studies made their underlying research data openly accessible.

In this study, we propose that the negative impact of issues in front of open research data adoption can be tackled by the right institutional and infrastructural instruments. We define infrastructural instruments as the combination of technical elements (e.g., open data portals, (meta)data standards and formats and tools for processing, searching, analyzing, and visualizing data) as well as governance elements (e.g., mechanisms to enhance privacy and trust and interaction with other data providers and users) underlying open data sharing and use. Institutional instruments are defined as the combination of formal structures (e.g., policies, processes), informal structures (e.g., norms, culture), and enforcement characteristics or operational mechanisms that institutions can put in place to incentivize open data sharing and use. Examining the potential of these instruments in promoting open research data practices is valuable since currently, many researchers do not receive sufficient institutional and infrastructural support for data sharing and reuse. Therefore, the objective of this study is to understand the roles that infrastructural and institutional instruments (or their combinations) can play in promoting open research data sharing and use behavior in the field of Epidemiology.

In this research, we employed qualitative research methods. We adopted three research approaches: the systematic literature review (SLR) research approach, the case study research approach, and the workshop research methodology. First, to understand which infrastructural and institutional instruments can be used to influence researchers to openly share their research data and to use openly available research data, we conducted a systematic review. We

conducted the review on the SCOPUS database, complemented by grey literature, to arrive at a list of instruments whose roles we further examined in the case study. Our systematic review showed that infrastructural instruments can range from providing powerful search engines that are sufficient for data search needs to the availability of data management tools. It showed that institutional instruments can range from providing separate funds for research data management to offering trainings on open science and research data management.

The case study that we conducted in this research examines how the instruments that we synthesized in our systematic review influence Epidemiology researchers in open research data practices. The case study information sources were ten Epidemiology researchers and a research data management consultant who work at various University Medical Centers (UMCs) and universities across the Netherlands. We complemented the qualitative interview data with an analysis of policy documents and web pages of these organizations. During the interviews, our objective was to understand whether the proposed instruments were available to the researchers**,** and the extent to which the proposed instruments (can) influence researchers' open research data sharing and reuse behavior. We systematically coded the qualitative interview data and operationalized the concepts before conducting the analysis.

Our analysis first uncovered the important characteristics of Epidemiology that influence the open research data practices. We found out that Epidemiology researchers may likely deal with large datasets from cohort studies, which makes data collection tough. We understood that clinical work puts extra time pressure on Epidemiologists. We realized that obtaining informed consent for open data sharing from patients in a clinical context is relatively harder. We illustrated that in Epidemiology, research agendas could be very flexible, which means that researchers may develop research questions even after data collection ends. We also realized how the General Data Protection Regulation (GDPR) could be inhibiting data sharing practices, and how data anonymization could be much less powerful contrary to our literature findings. We also found out that in this field, as time passes, data lose scientific relevance quickly. We also explained the prevalence of certain data sharing types such as data sharing by collaboration and one-on-one data sharing in the field, while also describing the lack of an open data sharing culture in this field.

The infrastructural instruments that are considered to be highly important for open data practices in our case are *easy-to-use repository interfaces, compatibility between different data infrastructures, availability of powerful search engines, availability of overarching registry of repositories, infrastructure's offering of metadata on data collection* and *the infrastructure's compatibility with the domain-specific privacy requirements.* For example, we found out that for Epidemiology researchers, being able to find detailed descriptions of how the data were collected is an important motivator for reusing open research data from data repositories. We understood that Epidemiology researchers are not satisfied with search engines on the open data repositories, and they expect to be able to use advanced search queries -similar to those on popular reference search engines like PubMed. We also demonstrated that the lack of an overarching registry that connects all the repositories in Epidemiology is considered to be a demotivator of open research data reuse.

The institutional instruments that are considered to be highly important for open data practices in our case are *data steward support, working with research data managers, providing support for legal aspects (i.e. privacy) relating to open data practices,* and *recognizing and rewarding open research data sharing contributions.* For example, we demonstrated that being able to work with research data managers would have a positive influence on researchers' motivation for open research data sharing in Epidemiology, because researchers do not have sufficient time to prepare and maintain datasets for open access. However, we understood that only researchers with larger projects get the chance to work with data managers, and in practice, data managers do not have time for making datasets open in the Epidemiology departments that we examined. We also realized that there could be communication issues between legal departments and Epidemiology researchers, since legal teams have the reputation of being too strict towards open data sharing applications. We also found evidence that Epidemiology researchers would be much more motivated toward open data practices if they were to believe that such efforts are sufficiently recognized and rewarded in their field.

To establish the extent to which our case study findings on infrastructural and institutional instruments can be applied to other research fields and the general scientific community, we held a workshop with nine participants who are either data stewards or research data officers working in different research fields at a Dutch research university. The workshop showed that many findings of our case study apply to other research fields. We realized common problems across research fields such as the low findability of research data on repositories, the lack of financial resources for research data management, and the lack of structured communication between researchers and legal teams. Providing standardized metadata on data platforms also seems to be a common (positive) factor for open data practices across different fields. We realized that many technical fields (such as geosciences) produce significant volumes of research data, which requires data archives to be able to cope with large-scale data fast and inexpensively. We found out that, although trainings on (open) data sharing are useful in other technical fields (such as Electrical Engineering), the effectiveness of institutional-level trainings should not be overestimated considering that each research project has unique needs in the context of data sharing. Furthermore, changing data sharing motivations in qualitative research could be much harder since in practice qualitative researchers generally have more difficulties in comprehending the (value of) data sharing concept.

Our study concludes that certain infrastructural and institutional instruments have the potential to enhance open research data practices in Epidemiology by addressing a variety of different legal, cultural, technical, and organizational issues that inhibit open research data sharing and reuse practices in this field. Many instruments are found to be out of reach of Epidemiology researchers despite having a huge potential for enhancing motivations. For instance, researchers do not have sufficient access to research data managers, search engines with satisfactory functionality, overarching registries, or reward systems for research data sharing contributions. We understand that institutional instruments in Epidemiology have the potential to support open research data adoption by reversing the lack of an open data sharing culture with the right incentivization approaches, by the provision of financing, and by actively supporting researchers via engagements with data stewards, research data managers, libraries

and data privacy officers. Infrastructural instruments also have the potential for supporting open research data adoption via increasing the findability and interoperability of the research data, and also via catalyzing researchers' interaction with data infrastructures -considering the high number of technical skills that are required during open data practices. For open research data sharing reuse practices in Epidemiology, institutional instruments in Epidemiology could be in a more vital position since tackling the lack of an open data sharing culture in Epidemiology is considered to be the most essential step toward behavior change. Nevertheless, our study also shows that many infrastructural and institutional instruments complement one another in practice, thus they should be combined to increase their effectiveness.

To strengthen the role of infrastructural instruments in promoting open research data adoption in Epidemiology, we propose several recommendations concerning different actors. We recommend infrastructure developers/providers to enhance the findability of research data by expanding features, to consider data dictionaries as obligatory elements of data sharing, and to invest in data infrastructures that tackle privacy concerns. We recommend university managements and policymakers to restructure the communication between legal teams and researchers, clarify the role of data stewards, clearly establish what is expected from researchers and other supporting professionals in terms of capabilities, consider the library's future role for ICT support, focus on building field-specific open science curricula that target not just PhD students but all researchers, re-frame the data ownership concept and focus on data controllership, incorporate open research data contributions as a key part of Reward and Recognition Programs, and separate the concept of open metadata sharing from open research data sharing in data sharing policies.

Although previous research has extensively examined the benefits, barriers, and motivators of open research data sharing and reuse, it has not looked at how different infrastructural and institutional instruments influence open research data sharing and reuse practices in a specific field. This study has academic relevance since it is, to our best knowledge, the first study that focused on the field of Epidemiology while examining the roles of instruments in open research data adoption based on field-dependent characteristics.

Tackling the barriers to open research data adoption benefits not only researchers and research communities, but also the society. Open research data adoption is valuable for public health because accessing data is considered to be a vital prerequisite for identifying public health problems that necessitate urgent responses. Moreover, because of increased transparency of the research processes and enhanced perception of scientific knowledge being a public good, the general public would build more trust in research. Open data adoption can also support the fundamental right of access to knowledge and lower inequalities due to the imbalance of research resources across the globe. Our study can inform governmental policymakers and lawmakers who want to tackle the barriers to data sharing stemming from the GDPR. University policymakers, funding agencies, and libraries can prioritize their interventions based on our study's indications of which of these tools are more promising than the others.

We recognize that there are limitations to this study concerning the research methods. For example, because we opted for the quota sampling approach in recruiting interviewees (i.e. recruiting interviewees from as many UMCs in the Netherlands as possible), we are not able to pose any conclusions about a specific type of Epidemiology researchers (e.g. PhD candidates or full professors) due to varying contextual aspects. Also, the researchers who participated in this study may or may not be representative of the Epidemiology field. Therefore, the results should not be immediately generalized to the wider Epidemiology field without replicating and validating the study by interviewing more people in the Epidemiology field. Furthermore, we recognize that there is always a possibility that some researchers may have given biased answers to our highly behavioral questions.

We recommend future research to conduct case studies in other contexts, considering that the issue that we examined in this research is a multi-actor issue and that in this study, we only focused on researchers. Examining the attitudes of policymakers in certain universities or examining the capabilities of infrastructure developers/providers could be valuable for understanding how the proposed instruments can be operationalized and realized in practice. Furthermore, we recommend researchers to replicate this case study with a different set of Epidemiologists to further understand the generalizability of the findings to the wider population of Epidemiologists.

Table of Contents

1. Introduction

Data can be viewed as the infrastructure of science (Tenopir et al., 2011). Due to advancements in computing and communication technologies, the amount of data collected, analyzed, and stored has increased tremendously, particularly in the last two decades (Tenopir et al., 2011). This highlights the shift to a new research paradigm, which could be named as "data-intensive scientific discovery", where all the findings of the science and the accompanying material exist on the internet and all interact with each other (Hey et al., 2009). In other words, science is making a paradigm shift towards data-driven research, where data intensity and collaboration have deemed research data sharing a necessity (Kurata et al., 2017; Tenopir et al., 2011).

The term research data refers to information that is used to validate research findings in research (*What Is Research Data?*, n.d.). Research data can be defined as "a well identified set of data that has been produced (collected, processed, analyzed, shared & disseminated) by a (again, well identified) research team. The data have been collected, processed and analyzed to produce a result published or disseminated in some article or scientific contribution." (Gomez-Diaz & Recio, 2022). The act of "sharing" data refers to the transaction of data from one actor (or organization) to another (*What Is Data Sharing?*, n.d.). Thus, research data sharing refers to providing access for the use and reuse of research data (Tenopir et al., 2011). Research data sharing may happen on a request of a researcher, which can be referred to as one-on-one research data sharing. On the other hand, such a data sharing can be done proactively by researchers, which relates to making research data "open" (i.e., enabling it to be fully discoverable and usable by others) (Burwell et al., 2013). Therefore, open research data can be defined as a structured set of data "that is actively published on the internet for public re-use, and that is freely accessible, usable, modifiable, and sharable by academic researchers" (Zuiderwijk & Spiers, 2019, p. 229). This research solely focuses on examining methods to increase openly sharing (or reusing) research data (i.e. researchers proactively sharing their research data with others). One-on-one research data sharing that happens on the request of a researcher is out of this study's objective.

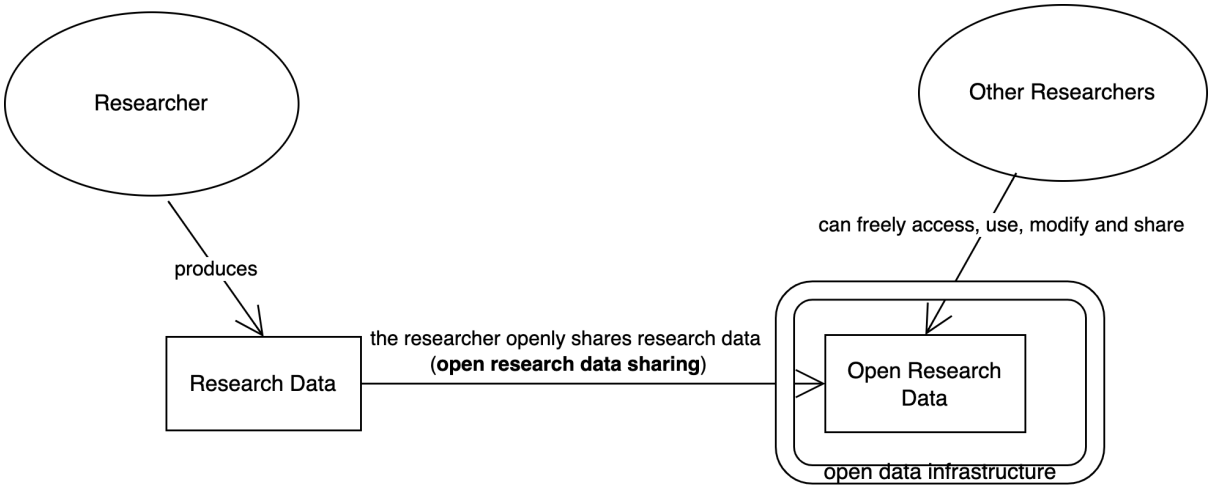

Figure 1 Open Research Data Sharing

1.1. State-of-art knowledge in open research data

This subsection explains important concepts and state-of-art knowledge concerning open research data sharing. These are the benefits of open research data sharing and reuse, researchers' motivations towards sharing and reusing open research data, field dependency in the open research data adoption, lack of open research data adoption in the field of Epidemiology, and finally, infrastructural and institutional instruments in the context of open research data.

1.1.1. Benefits of open research data sharing and reuse

Table 1 Benefits of open research data sharing and reuse

1. Facilitates application to different contexts (Patel, 2016).
2. Facilitates getting more citations (Patel, 2016).
3. Increased trust in the data (Tenopir et al., 2011).
4. Increased transparency in the research and data collection process (Patel, 2016).
5. Saves researcher time/effort (Tenopir et al., 2011).
6. Increased collaboration among researchers (Institute of Medicine, 2013).
7. Highlights research is a public good (Institute of Medicine, 2013).
8. Allows for metanalyses (Institute of Medicine, 2013).

The idea of research data sharing and reuse emerged due to the many benefits of data sharing. Some benefits of research data sharing and reuse are the following: (1) Data that are reused by other researchers can be applied in various other contexts easily (Patel, 2016). (2) Data collectors can get more citations by making their research data freely accessible and reusable (Patel, 2016). (3) When research data are reused by other researchers, the authenticity and objectivity of the data are confirmed, and the data are protected against misconduct regarding fabrication and falsification, therefore trust in the data increases (Tenopir et al., 2011) (4) Data sharing brings transparency to the research process and the data collection methods (Patel, 2016). (5) Data sharing enables researchers (who reuse research data) to save time, which means that they can use this time for other research activities (Tenopir et al., 2011). (6) Research data sharing and reuse strengthen collaborations among researchers (Institute of Medicine, 2013). (7) Research data sharing supports, strengthens, and honors the idea that research (scientific knowledge) is a public good (Institute of Medicine, 2013). (8) Research data sharing and reuse enable researchers to make meta-analyses by combining different data sets, therefore producing scientific knowledge that would not be possible to obtain without open research data (Institute of Medicine, 2013).

1.1.2. Motivations towards open research data sharing and reuse: drivers and inhibitors

Important funding agencies in the United States and Europe such as the National Science Foundation, Research Councils UK, and the European Commission have started to make data management plans mandatory under their grant applications (Kurata et al., 2017). Across the globe governments, funding agencies, universities and journals are also implementing data sharing policies that mandate research data sharing to different extents (Kurata et al., 2017).

Despite the paradigm shift and the apparent policy push from the governing institutions, it is not clear if data sharing is overall a prevalent practice throughout the entire scientific community (Kurata et al., 2017). Previous studies focused on understanding the factors positively or negatively influencing researchers to make research data open. One key publication, that of Tenopir et al. (2011), examines data sharing perceptions, barriers, and enablers of data sharing by conducting a survey. Results show that scientists choose not to share their data for many different reasons, ranging from insufficient time to lack of funding (Tenopir et al., 2011). Moreover, getting proper citations and the ability to learn about publications that use their data could be some of the key enablers of open research data adoption (Tenopir et al., 2011). The fact that correct data management plans are not established (or that they are not made mandatory) could also be an important barrier, as more than half of the participants in the survey said that their primary funding institution does not require them to do so (Tenopir et al., 2011). Again, nearly half of the participants said that their organization or project does not "provide the necessary funds to support data management during the life of a research project" (Tenopir et al., 2011, p. 5.). This could mean that data management support and policies are key determinants behind open research data adoption. Data standardization is also found to be an important factor, as many researchers find the current tools for preparing metadata lacking (Tenopir et al., 2011). There is evidence that researchers are generally positive about sharing data but many researchers do not receive sufficient institutional and infrastructural support for data sharing and reuse (Zuiderwijk, 2020).

On a different side of the discussion lie ethical concerns, which seem to be of higher importance in fields dealing with humans as study subjects. Pearce & Smith (2011) argue that in the field of Epidemiology, the adoption of open research data sharing may not be simple and straightforward because the ethical issues are "highly specific to each study, the nature of the data collected, who is requesting it, and what they intend to do with it" (Pearce & Smith, 2011, p. 1). They highlight the fact that in Epidemiology it is not always possible to fully hide the identity of study participants regardless of whether the data set is anonymized or not, which is a confidentiality issue (Pearce & Smith, 2011). Another key issue is the possibility of the open data easily being obtained by hostile agencies with vested interests in the outcome of the study (Pearce & Smith, 2011). The authors state, "for every independent epidemiologist studying the side effects of medicines and the hazardous effects of industrial chemicals, there are several other epidemiologists hired by industry to attack the research and to debunk it as 'junk science'." (Pearce & Smith, 2011, p. 5). When research data are freely accessible, companies can easily hire consultants to criticize the research publicly, even before publication (Pearce & Smith, 2011). Historically there have been many examples of such stigmatization of

unwelcome research findings, such as the industry efforts on influencing the studies on the toxicity of benzene and diesel particulate matter (Pearce & Smith, 2011). Therefore, Pearce & Smith (2011) defend the adoption of "restricted" access as opposed to "open" access. Restricted access involves sharing unaltered confidential data with external researchers while preserving confidentiality (Pearce & Smith, 2011). This can be done by giving access only to researchers with legitimate research questions or letting researchers access the data in a physically and electronically secure facility monitored by the data steward (Pearce & Smith, 2011).

Recognizing the massive diversity of factors behind open research data adoption, various authors examined the factors influencing open research data adoption and built theoretical frameworks to explain them. A key study is that of Zuiderwijk et al. (2020) who performed a systematic overview on the existing literature, synthesized and positioned the factors into eleven distinct categories, which are the following: the researcher's background, requirements and formal obligations, personal drivers and intrinsic motivations, facilitating conditions, trust, expected performance, social influence and affiliation, effort, the researcher's experience and skills, legislation and regulation, and data characteristics. Another example is the study of Sayogo & Pardo (2013), which provides a conceptual model explaining the likelihood of a researcher publishing their work openly via the following variables: (1) Social, Organizational, & Economic Related Challenges, (2) Legal & Policy Related Challenges, (3) Technology Related Challenges, (4) Local Context & Specificity Related Challenges.

1.1.3. Field dependency in researchers' behavior towards open research data sharing and reuse

Open research data sharing and reuse have become a common practice in certain research fields such as earth and planetary geophysics (Tenopir et al., 2018) and genetic research (Kurata et al., 2017). Indeed, there is evidence that the issues regarding researchers' adoption of open research data differ by the field (Kurata et al., 2017; Pearce & Smith, 2011; Zuiderwijk, 2020). Kurata et al. (2017) argue there is a complex and diverse relationship between data and research practices. This is because "data" itself is an umbrella term, and data exist in a specific context (and therefore it is a meaningless concept in isolation) (Kurata et al., 2017). Because research activities and data are intertwined, it would not make sense to evaluate data sharing independently from the research activities themselves (Kurata et al., 2017). Tenopir et al. (2018) argue that the fields that have more tendency towards open research data adoption are those that are not dealing with human subjects, those that have large large-scale instrumentation shared by many to collect data, those that have established metadata standards, and those that have a history of data sharing openness. Earth and planetary geophysics (Tenopir et al., 2018) and genetic research (Kurata et al., 2017) indeed could be qualifying for such characteristics. Tenopir et al. (2018) examined the practices and attitudes of researchers towards open research data sharing and reuse in the field of geophysics, a field that is widely known to have higher levels of data sharing. The authors showed that researchers in this field are concerned about the potential misuse of their data and they show the need for adequate citation and acknowledgment (Tenopir et al., 2018). Zuiderwijk & Spiers (2019) did a study on Astrophysics, another field with a known culture of open access to research data. The authors

showed that important factors that decrease Astrophysics researchers' motivation to openly share and reuse open research data are the enormous volume of some datasets and the lack of facilitating conditions (Zuiderwijk & Spiers, 2019). Moreover, the authors argue that in the Astrophysics field, more research data would be published if journals and research data centers could play a more proactive role (Zuiderwijk & Spiers, 2019).

1.1.4. Possible lack of data sharing in Epidemiology

Research shows that generally, researchers have become more willing to share the findings of their research data and use data of other researchers in the last decade (Tenopir et al., 2015). There is a global trend and consensus that research data should be shared, and this trend is being echoed in the decisions of major funding bodies, such as the EU Research Program Horizon 2020, which made research data sharing mandatory (Burgelman et al., 2019). However, in certain fields, data sharing behavior is still observed at very low levels. Epidemiology is one of the fields where research data sharing and use can be especially valuable yet more troublesome due to its nature. Epidemiology can be defined as "the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems" (Last, 2001, p.61). Thus, the field of Epidemiology is concerned about health-related phenomena in populations, and it is a method to find the causes of such occurrences (CDC, 2012). Sharing Epidemiological research data across different countries and continents can help understand the spread of diseases much faster. This benefit concerning speed is especially important for Epidemiology, because the field of Epidemiology often "races with time" when new diseases emerge.

The health emergencies caused by major outbreaks of Zika and Ebola had also already proved the necessity for open research data sharing in Epidemiology (Lucas-Dominguez et al., 2021). However, despite its apparent benefits, still, open research data sharing may not be a common practice in Epidemiology even after the emergence of the Covid-19 pandemic (Lucas-Dominguez et al., 2021). Research shows that despite the massive amount of Covid-19-related research published in the first five months of the pandemic, only 13.6% of these publications made their underlying research data openly accessible (Lucas-Dominguez et al., 2021). One case study done in the context of data sharing in the MERS outbreak shows that some barriers can be the lack of unified international standards for data handling, the existence of a time-consuming legal framework and authorization process, and the prioritization of scientific publication over data sharing (van Roode et al., 2018).

The fact that the field directly engages with humans as study subjects (and therefore deals with a lot of sensitive data) could be an underlying reason behind certain barriers to open research data adoption in Epidemiology. It also could be the source of justifications on why open data sharing does not, or should not, happen in this field. For example, Pearce & Smith (2011) draw attention to the ethical issues of giving open access to Epidemiological data: if open access was adopted in Epidemiology, then the informed consent process (an ethical and legal requirement in Epidemiological and medical research) should include asking permission for open data sharing (Pearce & Smith, 2011). The dilemma lies in the possibility that stating in an informed

consent process that the participant's data can be used "by anyone for any intention in the future" could easily result in people not wanting to participate in Epidemiological studies anymore (Pearce & Smith, 2011). There could also be other important reasons that may justify why open research data sharing does not take place in Epidemiology. Regardless, acknowledging the low levels of data sharing practices in Epidemiology as well as the various potential benefits that data sharing could provide to this field together imply the need to examine data sharing in the field of Epidemiology.

1.1.5. Infrastructural and institutional instruments and arrangements

Infrastructural and institutional instruments could be important tools to enhance open research data sharing and reuse. As there are various definitions for these concepts (data infrastructures and institutions) in the literature, it is necessary to provide one single definition for each concept for this study. One key definition for open data infrastructures is given by Zuiderwijk (2015), who combined the literature from digital infrastructures and the literature from information infrastructures to propose a definition for Open Governmental Data (OGD) infrastructures. Digital infrastructures can be perceived as a "collection of information technologies and systems that jointly produce a desired outcome" (Henfridsson & Bygstad, 2013, p. 6), and they may also encompass organizational structures and associated services, and facilities that are needed for the functioning of an industry or an enterprise (Tilson et al., 2010). On the other hand, Hanseth and Lyytinen (2010) define an information infrastructure as a "shared, open (and unbounded), heterogeneous and evolving socio-technical system (which we call installed base) consisting of a set of IT capabilities and their user, operations and design communities" (p. 4). Looking at these two pieces of literature, the definition of an open data infrastructure (and also an associated infrastructural instrument) should not only have technical but also social, organizational, and governance elements. Therefore, in this thesis project, infrastructural instruments are defined as the combination of technical elements (e.g., open data portals, (meta)data standards and formats and tools for processing, searching, analyzing and visualizing data) as well as governance elements (e.g., mechanisms to enhance privacy and trust and interaction with other data providers and users) underlying open data sharing and use. In the broadest sense, institutions can be defined as "rules" of the game in society (North, 1990). More specifically, institutions are rights, rule of law, and (political) constraints that shape human interaction and incentives in human exchange, possibly in a social, political, or economic sense (North, 1990; Williamson, 2009). Such rules are often enforced through different mechanisms, ranging from court rulings to societal pressure (Hodgson, 2006). There is also an important distinction between formal institutions (which are legal, written, and explicit) and informal institutions (which are informal, inexplicit, and unwritten) (Hodgson, 2006). Therefore, the following definition will be used for institutional instruments in this study: the combination of formal structures (e.g., policies, processes), informal structures (e.g., norms, culture), and enforcement characteristics or operational mechanisms that institutions can put in place to incentivize open data sharing and use. Finally, an "arrangement" will be defined as the combination of two or more instruments.

It shall be noted that such "factors", "drivers" and "inhibitors" evaluated and described in the literature of open research data are highly intertwined with the "institutional and infrastructural instruments" (which will be the focus of this thesis), since the majority of factors (e.g. formal obligations, facilitating conditions, legislation and regulation, social influence, data characteristics, etc.) are constructs that have been defined under specific institutional and/or infrastructural environments. For example, the factor "legislation and regulation" could refer to copyright and license issues, data policies, and national and international agreements (Zuiderwijk et al., 2020) which can all be considered "institutions" as they constitute either formal or informal rules. Therefore, many factors behind open research data adoption either directly refer to the specific institutional and/or infrastructural instruments, or they are highly intertwined with them.

Zuiderwijk (2020) states that the negative impact of challenges in front of open research data adoption can be mitigated by the right institutional and infrastructural arrangements. Therefore, examining the institutional and infrastructural arrangements that may support open research data sharing and reuse in this field is a crucial step toward promoting open research data sharing and reuse. For instance, examining the role of data management plans (as institutions) and the function of data repositories (as infrastructures) could be crucial to understand their potential for promoting research data sharing and reuse (Zuiderwijk, 2020). As such instruments could also differ across fields, it is necessary to investigate which instruments work well under which conditions in Epidemiology, and to what extent the instruments in Epidemiology may differ from the ones in other fields.

1.2. Literature gap and the research objective

Following the state-of-art knowledge documented above, the literature and knowledge gap is synthesized in this subsection. Subsequently, the bridge between this gap and the objective of the research is presented.

The three points of literature and knowledge gap could be presented as the following:

First, recognizing that factors behind open research data adoption differ heavily by field, field-specific studies are needed. The willingness of a researcher to share and reuse data depend a lot on the context (Zuiderwijk, 2020) and the practice of research data sharing is heterogeneous (Kurata et al., 2017). Although several studies on the motivations behind data sharing and reuse have been conducted, these studies do not provide in-depth insight into discipline-specific challenges and opportunities (Zuiderwijk & Spiers, 2019). This is also in line with authors of such studies (such as the study of Zuiderwijk et al. (2020)) recommending future research to empirically test the usability and completeness of their studies and to adapt them to specific contexts of open data sharing and use behavior. Studies conducted in the general research community cannot indicate the importance of drivers and inhibitors (Zuiderwijk et al., 2020). A discipline-specific examination is necessary to understand which drivers and inhibitors (or alternatively, which *institutional and infrastructural instruments*) are more important than others since such importance ordering would be a key input to open

research policy formation in individual fields. Such an analysis is crucial to understand the right institutional and infrastructural arrangements that may support the adoption and therefore reverse the low trends of open data sharing in individual fields, since copying arrangements across fields may not work (Zuiderwijk, 2020).

Second, considering the various potential benefits that open data sharing could provide, the field of Epidemiology is valuable to examine, as it currently has relatively lower levels of data sharing practices. Previous research suggests that in disciplines such as medicine (a close discipline to Epidemiology) or social sciences where human subjects or other restrictions may come into play, there may be less motivation toward open research data sharing and reuse (Tenopir et al., 2011). Especially during the Covid-19 pandemic researchers have had low tendencies toward research data sharing and reuse in the field of Epidemiology (Lucas-Dominguez et al., 2021). Although there are some studies on fields with higher open research data adoption rates, to the author's best knowledge, there is not a study examining the field of Epidemiology specifically. Considering that open research data sharing and reuse could bring many benefits to this field, Epidemiology is a field worthy of examination. Conducting a study on the field of Epidemiology now (during the second year of the Covid-19 pandemic) could especially be valuable since it can detect whether there have been any changes in motivations, perceptions, practices, or institutional/infrastructural environments as a result of the pandemic. Moreover, a study on Epidemiology would enable making comparisons with other fields where open research data adoption is higher, which could help understand why certain fields are doing better than the others in open research data adoption.

Third, there is evidence that many researchers do not receive sufficient institutional and infrastructural support for data sharing and reuse although researchers are generally positive about sharing da**ta** (Zuiderwijk, 2020)**.** This suggests the need to converge and focus specifically on infrastructural and institutional issues and possibilities rather than focusing on all the factors (including the ones that purely relate to intrinsic and personal motivations) behind the motivations of open research data sharing. Focusing on institutional and infrastructural arrangements also gives policymakers a better indication of their decision space and possibilities for change, because the factors that concern institutions and infrastructural instruments are the ones that are easier to influence with policy decisions.

This master thesis project, therefore, aims to understand what role infrastructural and institutional arrangements can play in promoting open research data sharing and use behavior in the field of Epidemiology.

The rest of this report structured as follows: Chapter 2 discusses the research approach that is adopted in this study. Chapter 3 discusses the infrastructural and institutional instruments that influence open research data adoption. The case study that is conducted in the field of Epidemiology is described in chapter 4. Chapter 5 explains the workshop where the case study findings are evaluated in the context of transferability. Chapter 6 provides recommendations to different actor groups and chapter 7 gives the conclusion to this master thesis report.

2. Research approach

This chapter starts by presenting the main research question and the research methods of this study. Then it explains the research activities and formulates the subquestions from the main research question, while also explaining the selection of methods, tools, and data needed for executing the research.

2.1. Main research question and research methods

In line with the research objective developed in the previous chapter, the following main research question is formulated: ***What roles can infrastructural and institutional arrangements play in promoting open research data sharing and use behavior in Epidemiology?***

To answer the main research question, the project employs qualitative research methods. This thesis will employ three research approaches. First, it will employ a systematic literature review (SLR), or in other words "a systematic review", research approach (Gopalakrishnan & Ganeshkumar, 2013). A systematic review can be defined as a summary of the literature that uses reproducible and explicit methods to systematically search, critically appraise, and synthesize on a specific issue (Gopalakrishnan & Ganeshkumar, 2013). Therefore, a systematic literature review identifies, selects, and critically appraises research to answer a formulated question (Dewey & Drahota, 2016). The benefits of systematic reviews include being able to comprehensively scan the literature on a specific topic (Green, 2005), and by executing a "fixed process", achieving transparency as well as replicability of the research (Mallett et al., 2012). Therefore, a systematic literature review should be transparent, clear, integrated, accessible, and focused (Pittway, 2008). Despite the aforementioned advantages, the systematic literature review research method also has several limitations. The first notable limitation is that conducting systematic reviews is time-intensive: a researcher has to systematically assess a high number of papers at the first stages of the search strategy (Mallett et al., 2012). Furthermore, another possible source of limitation is publication bias, which refers to researchers' tendency to favor publishing statistically significant data (Drucker et al., 2016). The publication bias affects systematic reviews because it results in the review favoring positive findings during study selection (Drucker et al., 2016).

The second research approach that this thesis project will adopt is the case study research methodology. The case study approach allows researchers to explore complex issues in-depth in their real-life settings (Crowe et al., 2011). Thus, this research approach is useful when there is a need to examine events or phenomena in their natural real-life context, especially when the boundaries between context and phenomenon are not as clear (Crowe et al., 2011; Yin, 2018). A paper that has similar objectives to this master thesis, the one of Zuiderwijk & Spiers (2019), also uses a case study research method to examine the motivations behind open research data sharing and reuse in the field of Astrophysics. The case study approach is suitable for studying researcher behavior and motivations in a specific field, as it allows investigating complex real-life events that necessitate thorough examination (Zuiderwijk & Spiers, 2019). A case study

could also be viewed as essential for this research since the underlying motivations of researchers towards open research data adoption are context-dependent (Kurata et al., 2017; Pearce & Smith, 2011; Zuiderwijk & Spiers, 2019). The case study will include interviews with Epidemiology researchers as well as research data management professionals, and analyses of policy documents and websites of organizations that are examined. Despite the aforementioned advantages relating to observing phenomena in-depth, the case study research method also has certain drawbacks. For example, because the case study concerns a certain group of people, it is not possible to be sure whether this group is representative of the larger population (Mcleod, 2019). Therefore, the case study findings cannot immediately be generalized to the wider population unless the study is further replicated and validated (Mcleod, 2019). Another limitation of the case study method is researcher bias, which refers to the possibility of the researcher's subjective opinions affecting the assessment of the data (Mcleod, 2019). Finally, two additional drawbacks with using interviews as the source of qualitative data in the case study approach are (1) that researchers may give biased or unrealistic answers to our highly behavioral questions, and (2) that conducting interviews is a highly time-consuming data collection approach (Mcleod, 2019).

The final research approach of this master thesis is the workshop research methodology. Research workshops can be defined as assembling a set of people with the goal of learning and generating the data that is required to reach the objective of the workshop (Ørngreen & Levinsen, 2017; Shamsuddin et al., 2021). Workshop as a qualitative research approach is a promising method since it allows for engagements among the participants as well as between the participants and the facilitator (Ahmed & Mohd Asraf, 2018). Such engagement can be in the form of constructive feedback and collaborative discussions (Ahmed & Mohd Asraf, 2018). Furthermore, workshops also enable the participants to interact and collaborate while learning about a topic, and such collaboratively shared learning experiences can provide valuable information that would not be possible to obtain from other research methods (Ahmed & Mohd Asraf, 2018; Ørngreen & Levinsen, 2017). For example, valuable data from this method can come from "participant interactions and artefacts produced by participants in carrying out group tasks" (Shamsuddin et al., 2021, p. 2). Despite the aforementioned advantages, there are also drawbacks to workshops. An important drawback is a possibility of some participants becoming passive and reluctant to participate due to feeling intimidated by the highly immersive collaborative environment (Ørngreen & Levinsen, 2017). Furthermore, another limitation is that the results of the workshop depend heavily on the facilitator's performance: how much valuable knowledge is elicited from a workshop depends on whether the facilitator manages to "create a good atmosphere, facilitate the sense of giving each other space, and be sensitive to verbal and nonverbal communication" (Ørngreen & Levinsen, 2017, p. 78). In addition, workshops require a large time commitment as they are highly collaborative, and it is also difficult to predict exactly how long the workshop activities will take.

2.2. Research activities and subquestions

The main research question presented is dissected into four subquestions. These subquestions are clustered in four distinct research activities. This research design can be seen visually in the research flow diagram in Figure 2.

2.2.1. Building a conceptual framework

This first research activity focuses on understanding infrastructural and institutional instruments that influence open research data adoption. The research method is a systematic literature review, complemented by analyses of gray literature, such as white papers, reports, guidelines, and websites. Inspired by Zuiderwijk et al. (2020) who used a systematic literature review approach to examine the factors influencing researchers' open research data adoption, this research will also follow a similar systematic review process to answer the first subquestion. The output is therefore a conceptual framework, which will be used as an input for the upcoming research activities that shift the focus specifically on the field of Epidemiology.

- SQ1: What infrastructural and institutional instruments influence researchers to openly share their research data and to use openly available research data?

2.2.2. Validation through the case study

The second research activity is a case study that applies the conceptual model in a single field (i.e., Epidemiology) in the form of a single case study. It tests the extent to which the framework established in the previous research activity applies to the field of Epidemiology, and also examines if some instruments have more influence than others in the field of Epidemiology. The data for the case study will come from interviews held with Epidemiology researchers and research data management professionals. Policy documents and websites of organizations will also be analyzed. The output will be the presentation of the case study analysis findings. The analysis that is done to answer subquestion 2 will delineate the infrastructural and institutional instruments in Epidemiology, their role, influence, and importance in open research data sharing and reuse practices in this field.

- SQ2: How do infrastructural and institutional instruments influence researchers in openly sharing their research data and in using openly available research data in the field of Epidemiology?

2.2.3. Usability of the research findings in other research disciplines

This research activity focuses on understanding the extent to which the case study findings can be relevant for other research disciplines than Epidemiology. Since the case study focuses solely on the field of Epidemiology, it is valuable to understand what could be learned from this study that concerns different fields. Such an examination can help expand the contributions

of this research by understanding the value of the instruments in other research disciplines that are close (or far) from the Epidemiology field, and by enabling the generalization of certain findings of the discipline-specific case study to the overall scientific community. The concept of determining if the results of a study are also applicable to other contexts is known as establishing transferability in qualitative research (Curtin & Fossey, 2007). Korstjens & Moser (2018) define transferability as "the degree to which the results of qualitative research can be transferred to other contexts or settings with other respondents." (p. 121). Establishing transferability also strengthens the trustworthiness of the study (Curtin & Fossey, 2007). The data needed to answer subquestion 3 will come from a workshop held with data stewards and research data officers who work in different research disciplines in a research organization. The case study findings will therefore be evaluated via the workshop research approach.

- SQ3: To what extent can the case study findings on infrastructural and institutional instruments be applied to other research fields and the general scientific community?

2.2.4. Recommendations

The last research activity is focused on providing a discussion and recommendation on the infrastructural instruments and arrangements that may support and promote open research data sharing in the field of Epidemiology. The main motivation is to discuss how the effectiveness of instruments and arrangements can be enhanced in this field, possibly by restructuring elements in certain instruments or by proposing new instruments to the field. The data to answer subquestion 4 will come from the findings of the previous research activities.

- SQ4: How can infrastructural and institutional arrangements in the field of Epidemiology be enhanced so that they are more effective in promoting open research data sharing and reuse?

2.3. Research flow diagram

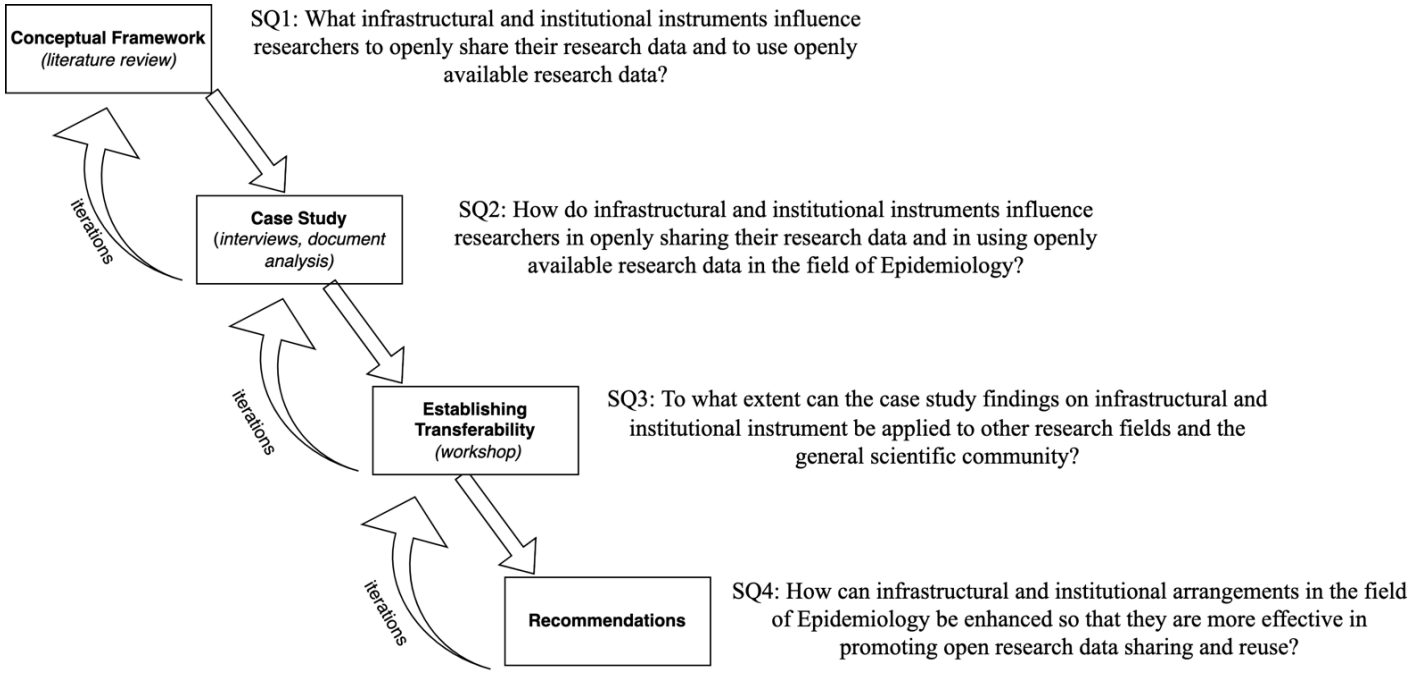

Figure 2 Research Flow Diagram

2.4. Deliverables of the research

The first deliverable of the research is a conceptual framework that explains the institutional and infrastructural instruments behind open research data adoption (chapter 3). The second deliverable is a case study in the field of Epidemiology, accompanied by not only interviews but also analyses of policy documents and websites of organizations (chapter 4). This case study intends to test this conceptual model to understand to what extent the conceptual model applies to this field and to evaluate whether certain instruments in Epidemiology are more important than others. This means an in-depth examination of the institutional and infrastructural instruments in the field. Infrastructural instruments can range from providing powerful search engines that are sufficient for open data search needs to the availability of data management tools. Institutional instruments can range from providing separate funds for research data management to trainings on open science and open data management. The third deliverable is the evaluation workshop, the main purpose of which is to understand the transferability of the case study findings (chapter 5). Based on our findings from the case study and its evaluation (workshop), the third deliverable is a recommendation on how institutional and infrastructural arrangements in Epidemiology can be enhanced to increase their effectiveness (chapter 6).

3. Infrastructural and institutional instruments in relation to open research data adoption

This chapter builds a conceptual framework on infrastructural and institutional instruments that influence open research data adoption. To build this framework, this chapter first presents two underlying theories (i.e. the institutional theory and technology acceptance models) in chapter 3.1. These theories will be used to support the analysis of the systematic review that will be subsequently conducted (in chapter 3.2.). Finally, the conceptual framework which is taken as a basis for the case study is presented at the end of this chapter.

3.1. Underlying theories and theoretical foundation for research

3.1.1. Selection of theories for the study

Theories function as lenses with which researchers can evaluate complicated issues, understanding where to specifically divert their attention when they are looking at certain data or providing a framework that will guide their analysis (Reeves et al., 2008). Using theories helps to develop a complex and comprehensive understanding of things such as how social groups and organizations work, operate, and interact with each other (Reeves et al., 2008). Such an understanding is valuable since research gains value only if the system that is examined is understood in depth.

In this research, the objective of using theories is to understand the working "mechanism" of the instruments that we will examine in the following sections (Figure 3). What we mean by "understanding the mechanism" of an instrument is establishing how this instrument contributes to open research data sharing and reuse behavior, or in other words, how it affects the factors (barriers or motivations) behind these behaviors. Zuiderwijk et al. (2020) and Zuiderwijk & Spiers (2019) argue that the factors influencing open research data sharing and reuse motivations are highly diverse. In line with this, understanding the mechanism by which instruments affect behavior is necessary for organizing instruments based on common characteristics in the form of a conceptual framework, and also for systematically evaluating their roles in the following chapters of this report.

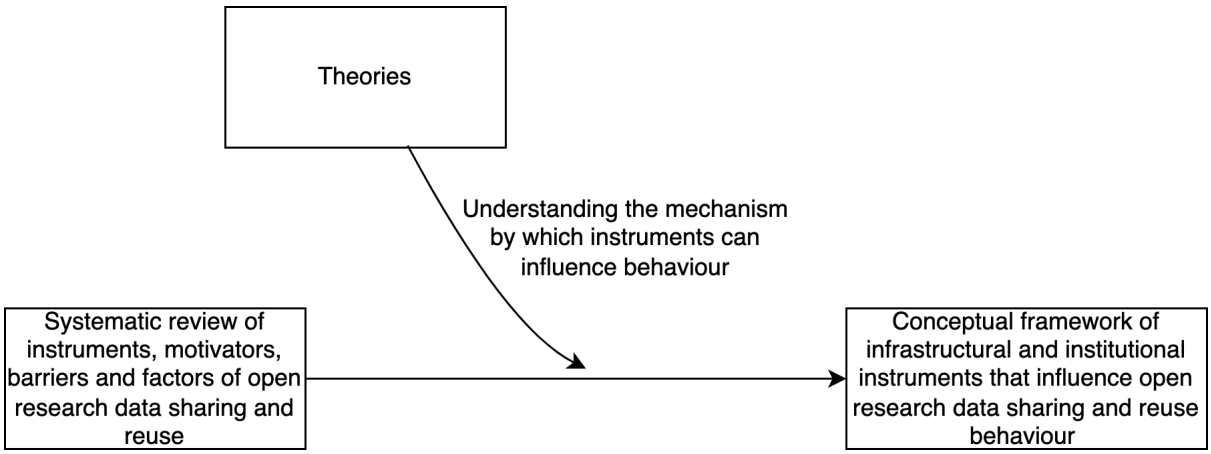

Figure 3 Use of theories in the study

Since this research project is in the field of open data, it would be intuitive to use theories directly from this field itself. However, there is limited theory development or theory extension in the field of open data, particularly regarding open data sharing and reuse (Zuiderwijk et al., 2020). That is why it is not possible to just use or extend theory from the field of open data. Furthermore, it seems that in the open data literature, few theories from other fields have been used, applied, or tested (Zuiderwijk et al., 2020). As open data (sharing and reuse) is a multifaceted construct, theories from other related disciplines such as those from information systems or psychology could be beneficial because theories from such fields also provide explanations for open research data adoption (Zuiderwijk et al., 2020). Acknowledging the limited theory development in open data literature, to complement this study with theory; it is decided that theories from other fields are used. The concept of using a theory from another field is sometimes referred to as "theory borrowing", which is using ideas from one domain to explain phenomena or constructs in other domains or a specific target case (Floyd, 2009). Although it is advantageous to use many different theories, considering the scope and limits of this study, one theory for institutional instruments and one theory for infrastructural instruments are to be chosen.

According to Grant & Osanloo (2015), an important step in selecting the most appropriate theory for a study is understanding "how the theory connects to your problem, the study's purpose, significance, and design" (p. 19). As we explained before, the theories connect to this study's design by establishing the mechanism concerning how an instrument affects the behavior of a researcher in the open research data sharing and reuse practice. Therefore, the first and essential criterion for an appropriate theory for our study is its consideration of "human behavior" as a central element in it. Furthermore, considering the definition of infrastructural instruments (see chapter 1.1.5.), an additional criterion of an appropriate theory for infrastructural instruments is its explicit consideration of technical artifacts (or technology) as a central element. Considering the definition of institutional instruments (see chapter 1.1.5.), an additional criterion for institutional instruments is its consideration of humans in social groups, that is, considering the interactions and engagements of humans with one another when explaining their behavior. Considering these criteria, a suitable theory for infrastructural instruments is technology acceptance models (TAMs) (which take their core from the theory of reasoned action (TRA) and theory of planned behavior (TPB)), since the models examine human behavior in the context of potential acceptance (or rejection) of a specific technology (Davis, 1986; Marangunić & Granić, 2015; Sharp, 2006). For institutional instruments, the institutional theory fits our criteria since it examines how social behavior is guided by distinct structural elements such as rules and norms (Scott, 2005). Furthermore, when the open data literature is examined, the suitability of these choices can also be confirmed: in their systematic literature review in the field of open data, Zuiderwijk et al. (2020) found that the theories that are mentioned most frequently are theory of planned behavior, the institutional theory, and technology adoption models.

Therefore, technology acceptance models are chosen for infrastructural instruments. Similarly, for institutional instruments, the institutional theory is chosen to be examined. Chapter 3.1.2.

and chapter 3.1.3. present the analysis of the literature review that is done to understand these theories in depth.

3.1.2. Theory of reasoned action (TRA), theory of planned behavior (TPB), and technology acceptance models (TAMs)

The technology acceptance model is a widely accepted model which is used to understand the human behavior toward potential acceptance (or rejection) of a technology (Davis, 1986; Marangunić & Granić, 2015; Sharp, 2006). The technology acceptance model takes its origins from the theory of reasoned action (TRA) and the theory of planned behavior (TPB), both of which are theories that emerged in the field of psychology (Marangunić & Granić, 2015).

3.1.2.1. Theory of reasoned action (TRA)

The theory of reasoned action argues that a person's behavior could be determined by evaluating their intention along with beliefs that the person would have for the given behavior (Marangunić & Granić, 2015). A person's behavioral intention is a function of two basic determinants, which are attitudes and subjective norms (Ajzen, 1985). It is argued that behavioral intention predicts behavior (direct effect) and that attitude only indirectly influences behavior via behavioral intention (mediated through intention) (Marangunić & Granić, 2015). The theory of reasoned action received criticism mainly due to the idea that people had little power over their behavior and attitudes. In that sense, the model is unable to deal with behaviors over which individuals have incomplete conscious control (Ajzen, 1985; Marangunić & Granić, 2015). This led to the development of the theory of planned behavior, which is just the addition of the "Perceived Behavioral Control" construct to the theory of reasoned action. "Behavioral Control" refers to the ability of that person in executing that behavior and the "Perceived Behavioral Control" refers to the perception of this: a person's confidence in their ability to perform that behavior (Ajzen, 1991), or in other words, a person's perception of the ease or difficulty of executing the behavior (Ajzen, 1985).

3.1.2.2. Theory of planned behavior (TPB)

The theory of planned behavior (Figure 4) argues that an individual's performance of a specific behavior is determined by their intention of performing that behavior (Marangunić & Granić, 2015). Intentions encompass the motivational factors that affect behavior (Ajzen, 1991). They imply how much effort a given individual is planning to make to execute the behavior, or in other words, how motivated the person is (Ajzen, 1991). The intention is determined by three constructs: (1) individual's attitudes toward the behavior ("Attitude toward the behavior"), (2) subjective norms about engaging with the behavior ("Subjective Norm"), and (3) perceptions of whether the person can successfully engage with the target behavior ("Perceived Behavioral Control") (Marangunić & Granić, 2015). Perhaps the most crucial element in this theory is "Perceived Behavioral Control" and its relation to "Intention" and "Behavior". The direct link from "Perceived Behavioral Control" to "Behavior" is justified by Ajzen (1991) via two

rationales: (1) looking at two people, if the intention is held constant, the person with higher confidence (i.e. perceived behavioral control) in performing a behavior will persist more than a person with lower confidence. Ajzen (1991) states "…even if two individuals have equally strong intentions to learn to ski, and both try to do so, the person who is confident that he can master this activity is more likely to persevere than is the person who doubts his ability" (p. 184). (2) The direct link from "Perceived Behavioral Control" to "Behavior" also could work to some extent as a substitute of "actual" control, which refers to nonmotivational factors as availability of opportunities and resources (Ajzen, 1991). Overall, the theory of planned behavior implies "To the extent that a person has the required opportunities and resources, and intends to perform the behavior, he or she should succeed in doing so." (Ajzen, 1991, p. 182).

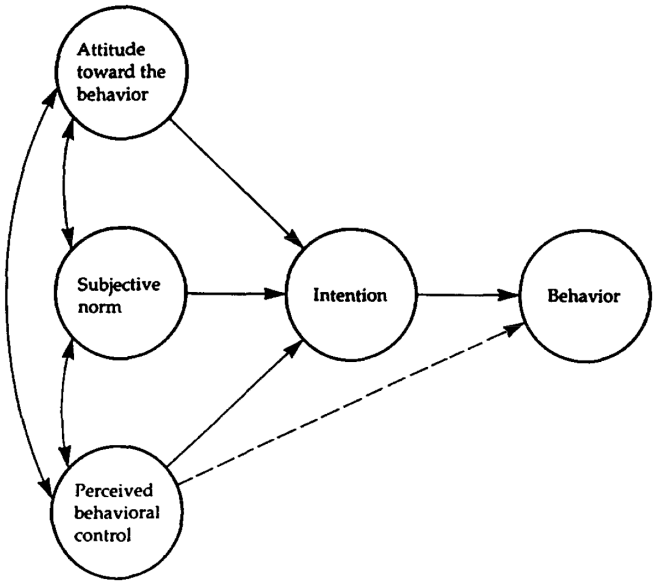

Figure 4 Theory of planned behavior (Ajzen, 1991)

3.1.2.3. Technology acceptance models: TAM 1 and TAM 2

The technology acceptance model (TAM) (Figure 5) emerged due to the need to find a reliable measure that could explain why a given system could face acceptance and rejection (Marangunić & Granić, 2015). Davis (1986) used TRA and TPB to come up with a reliable model that could predict the actual use of any specific technology (Marangunić & Granić, 2015). The main argument behind using these theories was that the actual use of a system could be viewed as a "behavior" (Marangunić & Granić, 2015). He identified two types of constructs "Beliefs" that predict the attitude of a system user towards using that system: "Perceived Usefulness" and "Perceived ease of use" (Davis, 1986). In TAM, a user's motivation is determined by three factors: "Perceived Ease of Use", "Perceived Usefulness", and "Attitude" toward using the system (Marangunić & Granić, 2015). Perceived usefulness refers to the degree to which an individual thinks that using that system will increase their job performance, and the perceived ease of use refers to the level of easiness that an individual believes they will experience while using the system (Sharp, 2006). Furthermore, system design characteristics (denoted by X1, X2, and X3 in Figure 5) directly influence "Perceived Usefulness" and "Perceived Ease of Use" (Marangunić & Granić, 2015).

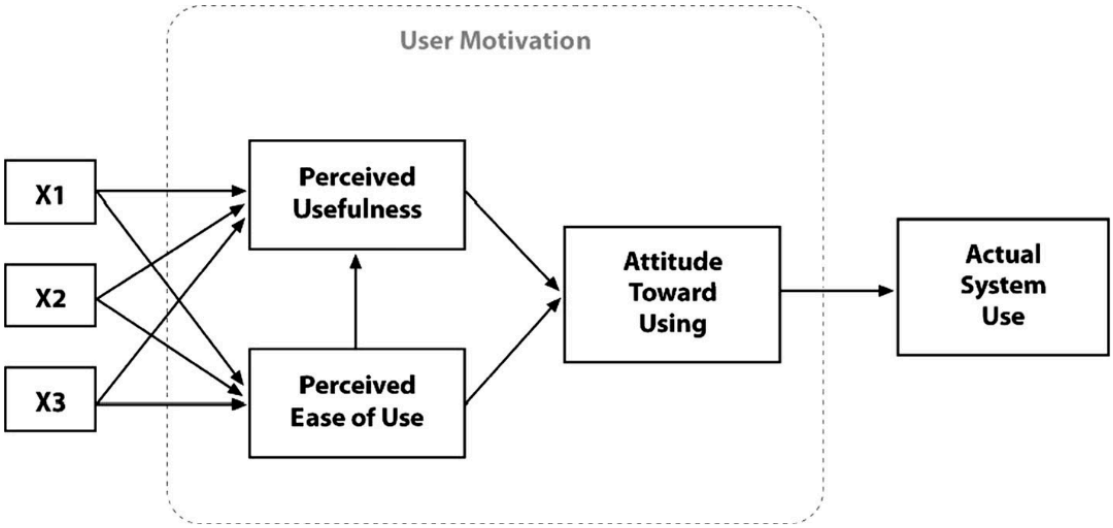

Figure 5 Technology acceptance model by Davis (1986)

Further modifications of the technology acceptance model have resulted in the replacement of the "Attitude" construct with the "Behavioral Intention" construct since a system that is perceived as useful by an individual could cause that individual to form a strong behavioral intention to use it without formation of any attitude (Marangunić & Granić, 2015). Furthermore, modifications were brought to further consider external variables that may influence the constructs concerning "Beliefs" (i.e., perceived usefulness and perceived ease of use). Venkatesh & Davis (1996) exemplified such external variables as system characteristics, training, user involvement in the design, and the nature of the implementation process. Upon further refinements of the TAM model, Venkatesh & Davis (2000) proposed the **TAM 2** model (Figure 6), which essentially aims to explain factors that explain perceived usefulness. These factors include "**Subjective Norm"** (the influence of others whom the system user considers as important on the system user's decision to make use of the technology), "**Image"** (the desire of a system user to maintain a favorable image within a reference group), "**Job Relevance"** (individual's perception of the degree to which the system applies to their job), "**Output Quality"** (considerations of what tasks a system can perform and the degree that those tasks match their goals) and "**Result Demonstrability"** (producing tangible results using the innovation) (Venkatesh & Davis, 2000). Furthermore, "**Experience"** and "**Voluntariness"** (the extent to which the user perceives that the decision to use the system is non-mandatory) are added to the model as moderating variables of the "Subjective Norm" (Marangunić & Granić, 2015; Venkatesh & Davis, 2000). In a nutshell, the model implies that both cognitive instrumental processes (i.e. job relevance, output quality, result demonstrability, and perceived ease of use) and social influence processes (i.e. subjective norm, voluntariness, and image) have a strong influence on a user's acceptance of a system (Venkatesh & Davis, 2000).

Technology acceptance models continued to be modified and augmented since the 2000s. One of the key important augmentations could be models adding the construct of trust, which is argued to be a variable influencing intention to use in several studies (e.g. Ghazizadeh et al. (2012); Ha et al. (2019)).

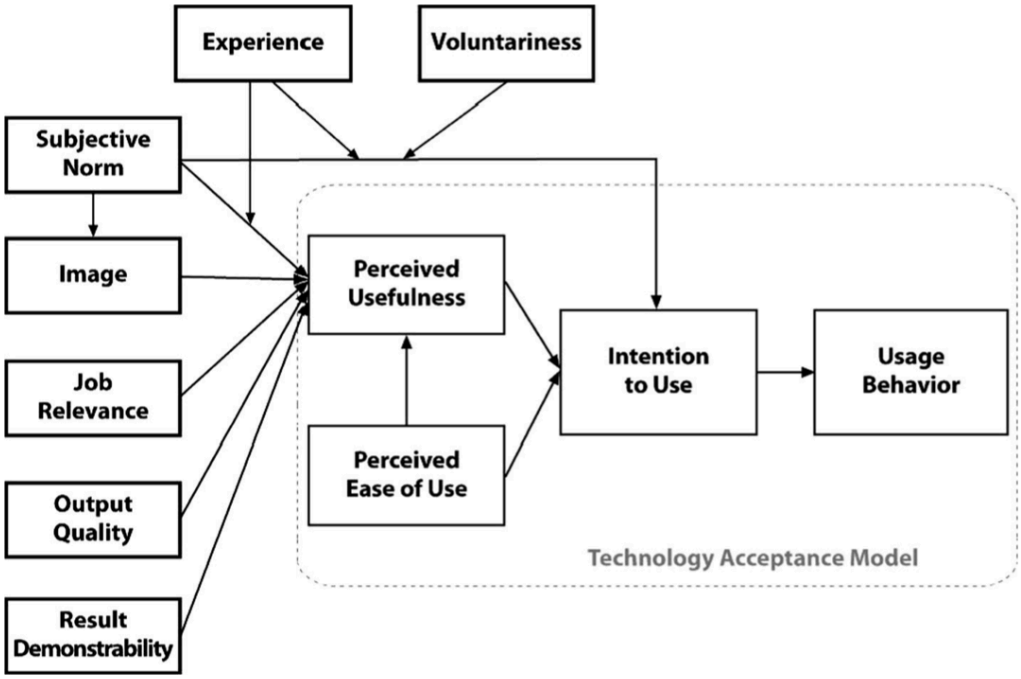

Figure 6 TAM 2 by Venkatesh and Davis (2000)

3.1.2.4. Suitability of the theory for the context of this research

The theory of reasoned action (TRA), the theory of planned behavior (TPB), and technology acceptance models (TAMs) all examine why people "accept" an innovation and start to use such a technology in their routine. An open data infrastructure can be an example of such a technology. Open data infrastructures are in fact infrastructures that emerged due to the internet becoming the essential infrastructure of the 21st century (*Delivering Digital Infrastructure Advancing the Internet Economy*, 2014). In this regard, open data infrastructures are novel. For many researchers who wish to share research data or reuse shared data, using an open data infrastructure could mean understanding a new system, learning how to use it appropriately, and therefore making the system a part of their new routine. Depending on the age of researchers, there could be many researchers who have not used a complex infrastructure like this up until a late point in their career since such infrastructures became available for use in the 21st century. This would deem viewing an open data infrastructure as a novel technology a necessity. If one views the infrastructures concerning open research data sharing and reuse as new technology under a sociotechnical system, it is, therefore, possible to examine why researchers do or do not adopt this technology using the theory of reasoned action (TRA), the theory of planned behavior (TPB) and technology acceptance models (TAMs). Therefore, these theories will be used to conceptualize the literature findings for the infrastructural instruments.

3.1.3. The institutional theory

Institutional theory is concerned with deeper and resilient aspects of social structure (Scott, 2005). It evaluates "the processes by which structures, including schemas, rules, norms, and routines, become established as authoritative guidelines for social behavior" (Scott, 2005, p. 2). It investigates how these elements are founded, spread, adopted, and modified over space and time, and how their use decline and after all, get discontinued (Scott, 2005). Institutions constitute regulative, normative, and cultural-cognitive elements that, together with associated activities and resources, deliver stability and meaning to social life (Scott, 2013). Therefore, bringing stability and order to social structures is a distinct feature of institutions (Scott, 2005, 2013). Institutions aim to establish the difference between acceptable and unacceptable behavior via imposing restrictions, which are created by defining legal, moral, and cultural boundaries (Scott, 2013). Putting prohibitions and constraints on the action also confirms an institution's capacity to control or constrain a given behavior (Scott, 2013). On the other hand, institutions also support and enhance the actions of actors via the provision of stimulus, guidelines, and resources for actions (Scott, 2013).

3.1.3.1. The three pillars of institutions

A key element in institutional theory is the three pillars of institutions: regulative, normative, and cultural-cognitive pillars (Table 2) (Altayar, 2018). Each of the "systems" that these pillars refer to (i.e. regulative systems, normative systems, cultural-cognitive systems) is a vital ingredient of institutions (Altayar, 2018; Scott, 2013). The regulative pillar highlights the regulatory processes, which are composed of **rule-setting, monitoring, and sanctioning activities**. (Altayar, 2018; Scott, 2013). To influence future behavior, -under regulatory processes- rules are established, the conformity of actors to such rules is evaluated, and -if needed- sanctions are modified and reimposed (Scott, 2013). Sanctions refer to rewards or punishments, they may involve informal mechanisms like shaming as well as formal mechanisms such as actions taken by courts under law. Often via **legal sanctions** which have **coercive mechanisms,** regulatory processes form the basis of legitimacy (Scott, 2013). Different theorists in different fields may view institutions as resting mainly on a specific pillar. The field of institutional economics often views institutions as resting only on the regulatory pillar, which is why institutional economists have often defined institutions as "rules" in society (North, 1990).

Table 2 Three pillars of institutions (Scott, 2013)

	Regulative	*Normative*	*Cultural-Cognitive*
Basis of compliance	Expedience	Social obligation	Taken-for-grantedness Shared understanding
Basis of order	Regulative rules	Binding expectations	Constitutive schema
Mechanisms	Coercive	Normative	Mimetic
Logic	Instrumentality	Appropriateness	Orthodoxy
Indicators	Rules Laws Sanctions	Certification Accreditation	Common beliefs Shared logics of action Isomorphism
Affect	Fear Guilt/ Innocence	Shame/Honor	Certainty/Confusion
Basis of legitimacy	Legally sanctioned	Morally governed	Comprehensible Recognizable Culturally supported

The normative pillar is concerned with social obligation and normative rules concerning adopting new social structures (Altayar, 2018). In that regard, normative rules are concerned with creating **social obligations**. The normative pillar is located under normative systems, which include two elements: (1) values, which refer to conceptualizations of what is "preferred" or "desired", as well as standards that are used to assess existing structures or behaviors concerning these conceptualizations; (2) norms, which refer to descriptions of how things should be done, or in other words, prescription of legitimate means in pursuing things that are valued (Scott, 2013). Attributing values and norms to only a selected set of people (instead of all members of the social collectivity) gives rise to the concept of **roles,** which refer to "conceptions of appropriate goals and activities for particular individuals or specified social positions" (Scott, 2013, p. 64). Roles can be constructed via formal means (e.g. in an organization, particular positions are defined to carry certain responsibilities and to give varying levels of power in accessing organizational resources) or they can emerge informally through interactions (Scott, 2013). Through beliefs, norms, and roles, the normative pillar is usually assumed to have the function of putting constraints on social behavior: "Given this situation, and my role within it, what is the appropriate behavior for me to carry out?" (Scott, 2013, p. 65). Scott (2013) argues that this indeed is why the normative conception of institutions is often embraced by sociologists, who, after all, examine institutions such as religious groups, communities, and social classes that give existence to values and norms.

The third pillar is the cultural cognitive pillar which focuses on the shared conceptions that form the nature of social reality and how meanings are constructed and created (Altayar, 2018; Scott, 2013). This pillar focuses on the cognitive cultural dimension of social constructs, it is argued that between the external stimuli and the response of individuals lie internalized symbolic representations of the world (Scott, 2013). Thus, the emphasis is on symbols and meanings. Symbols (words, signs, gestures) influence the meanings humans attribute to objects

and activities. The cultural cognitive pillar assumes that cultural conceptions vary from person to person: "Persons in the same situation can perceive the situation quite differently—in terms of both what is and what ought to be" (Scott, 2013 p. 68). Most importantly, in this pillar, compliance happens because people do not know the other "ways" or "options" they could have followed (i.e., other ways are inconceivable). **They follow routines because they believe that "way" is just "the way" they do these things**, which refers to orthodoxy being the logic in justifying conformity in this pillar (Scott, 2013). Cultural beliefs are often contested in times of social change (Scott, 2013).

3.1.3.2. Institutional pressures

Institutional pressures are those that can affect organizations and institutions (Altayar, 2018). Such institutional pressures have been discussed by Dimaggio & Powell (1983) under the context of institutional isomorphism (which refers to the study of processes making institutions converge to one another, or in other words, become "similar") (Dimaggio & Powell, 1983). There are three types of institutional pressures: coercive pressure, mimetic pressure, and normative pressure (Altayar, 2018; Dimaggio & Powell, 1983). Coercive pressure occurs due to **political influence and legitimacy**: an organization (X) (who is dependent on (Y)) gets pressured by organization Y (Dimaggio & Powell, 1983). Such pressures can take the form of force, persuasion, or invitation (Dimaggio & Powell, 1983). An example of this is when the government puts out new environmental regulations and mandates manufacturers to adopt new pollution technologies to comply (Dimaggio & Powell, 1983). Mimetic pressure occurs in times of uncertainty (Dimaggio & Powell, 1983). When the environment, the relevant technologies in it, or the organization's goals are uncertain or ambiguous, this puts mimetic pressure on an organization that goes through a change to be like another organization (Dimaggio & Powell, 1983). This mimicking is referred to as "modeling", and an example of such modeling is when new and successful governmental initiatives on one side of the world get copied by those on other sides of the world (Dimaggio & Powell, 1983). Normative pressure stems from professionalization, which is defined as the collective struggle that members of an occupation have in defining the methods and conditions for their work, and in establishing a cognitive base for their occupational autonomy (Altayar, 2018; Dimaggio & Powell, 1983). Such pressure can come from two streams: (1) formal education which is produced by an educational institution, and (2) growth of professional networks (Dimaggio & Powell, 1983). For example, universities and other education institutions (formal education) can develop organizational norms within a group of staff in an organization, or professional associations (professional networks) can produce normative rules about professional behavior (Dimaggio & Powell, 1983).

3.1.3.3. Rationalized myths

Another important element in institutional theory is rationalized myths, which could have the function of forming or altering organizational structure (Meyer & Rowan, 1977). The argument behind this function is that, when rational myths are produced and they manage to become common across different networks, they exert some sort of institutional pressure on

organizations and lead to change (Altayar, 2018; Meyer & Rowan, 1977). Altayar (2018) exemplifies that the adoption of Open Governmental Data (OGD) is due to the rationalized myths that preceded it. Before its adoption, OGD was presented with so-called "idealized views", which were simply the rationalized benefits of the system. OGD innovation was described as a solution that could improve transparency, give access to government information, have financial value, and support the public administration processes; and such descriptions (myths) have benefited its adoption and led to the formation of OGD as an institution (Altayar, 2018).

3.1.3.4. Suitability of the theory for the context of this research

The suitability of institutional theory to this research is related to our consideration of the issue of this study being a socio-technical issue (and not solely a technical issue). The issue of open research data sharing and reuse is a socio-technical issue because it is not only related to technical infrastructures on which research data sharing may be done, but also to aspects of social structures such as rules, norms, routines, legal contexts, and culture. Social structures shape the behavior of researchers and their motivations towards open research data sharing and reuse. The adoption of open research data is only possible if the related social system facilitates it and if there are proper guidelines for the behavior. It is no doubt that the act of sharing or reusing open data is the result of researchers' behavior towards the act. As the institutional theory examines how different social structures and elements may affect social behavior, one could make use of the institutional theory to understand how social structures should be (res)shaped and used, so that they become guidelines for the behaviors of open research data sharing and reuse (Scott, 2005). Therefore, the institutional theory will be used to conceptualize the literature findings for the institutional instruments.

3.2. Infrastructural and institutional instruments: systematic literature review

To build a conceptual framework on infrastructural instruments and institutional instruments that influence open research data sharing and reuse (behavior), it is important to understand which instruments can be used for open research data adoption. Furthermore, it is important to build the relation of these instruments with the various **barriers, motivators,** and **factors** of open research data practices. Therefore, to synthesize the (functions of) instruments from the literature, a systematic review is conducted in this chapter. Furthermore, we examine the mechanisms by which these instruments can influence open research data sharing and reuse behavior by using the theories we established previously. In chapter 3.2.2., we illustrate how we make use of theories in italics. Subsequently, we use the information that is gathered in this chapter to finally build the conceptual framework in chapter 3.3.

3.2.1. Study selection and assessment

In the identification phase, the search was conducted on the SCOPUS database, and only under English language papers, with the following 2 SCOPUS queries:

(1) "open research data" AND ("institutional" OR "institution*" OR "infrastructural" OR "infrastructure*") AND ("instrument*" OR "arrangement*") AND (LIMIT-TO (LANGUAGE , "English"))

(2)[1] "open research data" AND ("sharing" OR "share" OR "reuse" OR "use") AND ("factor*" OR "motivation*" OR "barrier*" OR "influence*") AND (LIMIT-TO (LANGUAGE , "English"))

In the identification phase, an in-press paper from Zuiderwijk & van Gend (in press) and its accompanying keynote speech by Zuiderwijk (2020) are also added to the list of literature since this master thesis project builds on the authors' existing work. Figure 7 presents the full search strategy and Table 3 presents the overview of the literature that is included in this systematic review.

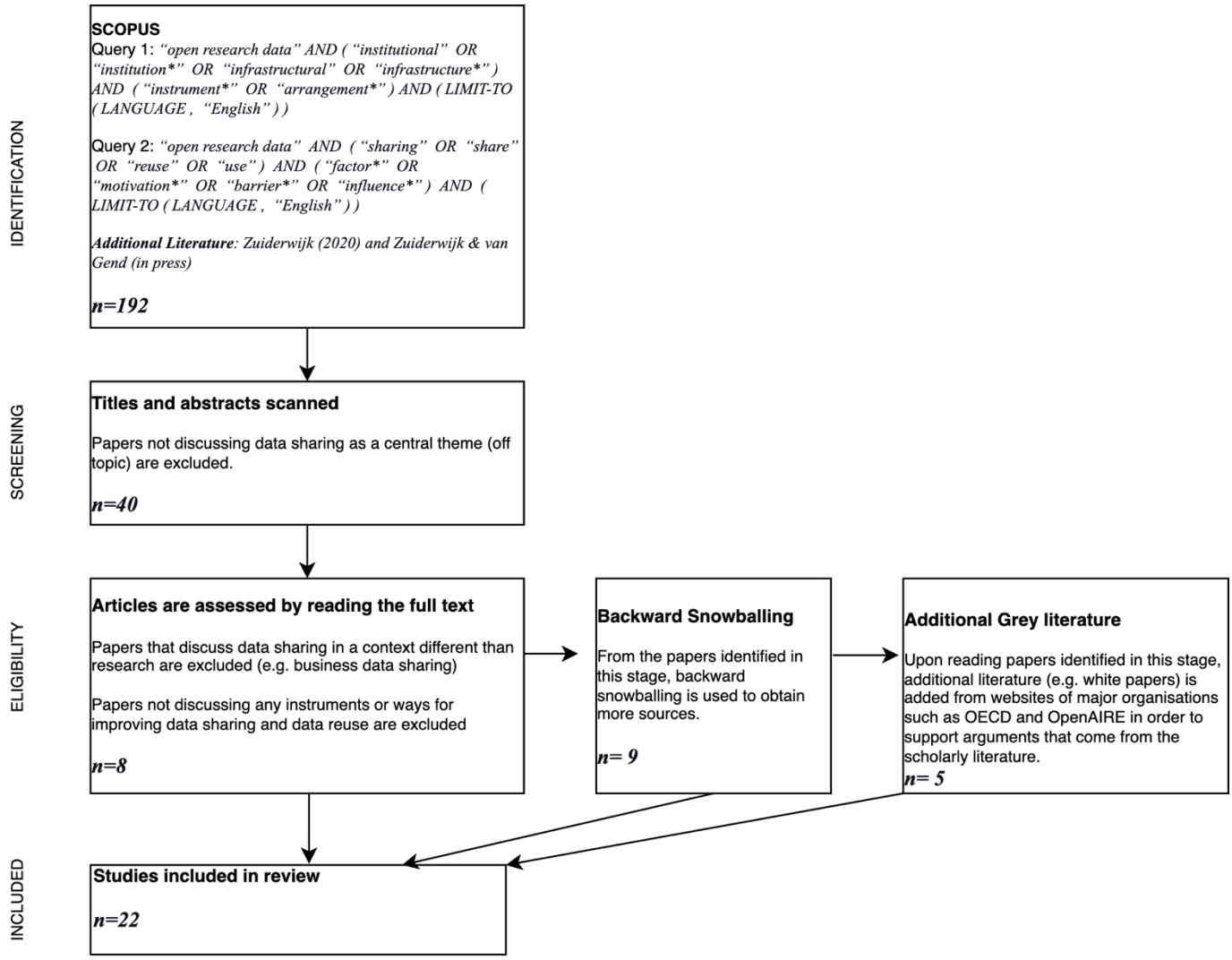

Figure 7 Search strategy for the literature review

[1] The purpose of including the second query, which does not include any keywords for institutional or infrastructural instruments, is the following: It is expected that simply searching the literature under the keywords "institutions" or "infrastructures" may not give satisfactory results for this study as these terms have varying definitions and usage across literature and contexts. These words relate to the "factors" influencing open research data adoption. This is why a second query was added to include alternative broader terms such as "factor", "motivation", "barrier" or "influence" to broaden the search scale.

Table 3 Overview of the sources included in the literature review

No	Authors	Title	Explanation
1	(Behnke et al., 2019)	Fostering FAIR Data Practices in Europe	The report discusses the features that data repositories should adopt in order to facilitate FAIR principles.
2	(Campbell, 2015)	Access to Scientific Data in the 21st Century: Rationale and Illustrative Usage Rights Review	The paper gives recommendations on desirable characteristics of open data repositories.
3	(Crosas, 2016)	Open Source Tools Facilitating Sharing/Protecting Privacy: Dataverse and DataTags	This webinar presents the Dataverse project and how the project specifically follows the FAIR data publishing standards.
4	(da Costa & Lima Leite, 2019)	Factors influencing research data communication on Zika virus: a grounded theory	This paper discusses factors influencing research data communication in the context of the Zika virus.
5	(Downs, 2021)	Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories	This article discusses the repository assessment and certification instruments that are used to improve repositories and to obtain the value intended by such data infrastructures.
6	(Fecher et al., 2015)	What drives academic data sharing?	This paper investigates the academic data sharing process in the eyes of the researcher who is sharing the data, and presents a framework that explains what elements go into this data sharing process.
7	(Harper & Kim, 2018)	Attitudinal, normative, and resource factors affecting psychologists' intentions to adopt an open data badge: An empirical analysis.	The paper investigates psychologists' intention toward obtaining an open data badge and discusses factors contributing to data sharing behaviors.
8	(Kim & Adler, 2015)	Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories	The paper discusses factors that contribute to data sharing behaviors among social scientists.
9	(Michener, 2015)	Ecological data sharing	This paper discusses a range of instruments that can be used to influence data sharing practices in the field of Ecology, touching upon the role of cyberinfrastructure and the role of policies.
10	(Neylon, 2017)	Building a Culture of Data Sharing: Policy Design and Implementation for Research Data Management in Development Research	The paper discusses a pilot project that aims to observe the implementation of data sharing and management requirements among research projects, and it shows whether and how the involved parties deal with the management of data.
11	OECD (2007)	OECD Principles and Guidelines for Access to Research Data from Public Funding	This guideline provides suggestions and information for governments and research institutions regarding how to overcome the challenges in front of research data sharing and access.
12	(Patel, 2016)	Research data management: a conceptual framework	The paper discusses problems in the context of research data management at the institutional level, and gives recommendations to organizations regarding how to better manage the research data cycle.
13	(Piwowar et al., 2007)	Sharing Detailed Research Data Is Associated with Increased Citation Rate	The paper discusses the correlation between sharing research data and getting more citations.
14	(Piwowar et al., 2008)	Towards a data sharing culture: Recommendations for leadership from academic health centers	In the context of sharing biomedical research and healthcare data, this paper provides recommendations to academic health centers for improving data sharing practices.
15	(Ringersma & Adamse, 2019)	Data Stewardship @ WUR: advice on a role for Data Stewards	This report gives guidance on what the data stewardship role involves, and how this role should be structured.
16	(Schmidt et al., 2016)	Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey	The paper discusses what is expected from infrastructures (functionalities) for sharing of data; and also, barriers and inhibitors to data sharing.
17	(Shelly & Jackson, 2018)	Research data management compliance: is there a bigger role for university libraries?	This paper presents possibilities for university libraries in supporting staff with research data management activities to make research data more accessible.

18	(Tenopir et al., 2012)	Current Practices and Plans for the Future	This white paper discusses the current practices of academic libraries regarding research data services support, and how these practices can be enhanced.
19	(Zuiderwijk, 2020)	Open Research Data sharing and use by means of infrastructural and institutional arrangements	This keynote presentation discusses infrastructural and institutional arrangements that can be used to enhance research data sharing and reuse.
20	(Zuiderwijk et al., 2020)	What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption	The paper presents a systematic literature review on the drivers and inhibitors behind researchers' motivation toward open research data practices.
21	(Zuiderwijk & van Gend, in press)	Open research data: a case study into institutional and infrastructural arrangements to stimulate open research data sharing and reuse	This paper discusses infrastructural and institutional arrangements that can be used to influence research data sharing and reuse in the form of a case study.
22	*Published in 2009 by the data initiative "PARADE"*	Strategy for a European Data Infrastructure	This white paper presents the features of a sustainable European data infrastructure and discusses the need for compatible data infrastructures.

3.2.2. Analysis of the systematic review

3.2.2.1. Infrastructural instruments influencing open research data sharing and reuse

The first infrastructural instrument category concerns the **usability of infrastructures** (e.g. data repositories and other complementing tools). The perceived effort to publish is a heavily indicated barrier in front of researchers' motivation for sharing and reusing openly available research data (Harper & Kim, 2018; Kim & Adler, 2015; Zuiderwijk et al., 2020), which deems instruments tackling this issue crucial. To properly store research data, there need to be suitable data repositories available to researchers, and such data repositories should be **able to accommodate large-scale data, should be easy to use, should enable an increase in data storage growth, and should be reliable** (Campbell, 2015; Kim & Adler, 2015; Zuiderwijk et al., 2020). *These instruments could ensure that an open data infrastructure is perceived as useful and easy to use by the user (the researcher), which is expected to increase researchers' attitudes toward using it, as TAM by* Davis (1986) *suggests. An infrastructure having the capacity for large-scale data and data storage growth essentially relates to the "output quality" construct affecting "perceived usefulness" in TAM 2: Researchers are more and more dealing with larger datasets, which means that their goal in using an open data infrastructure (output) is not just solely storing data, but storing large-scale data. Thus, the degree to which an infrastructure can perform the associated goals of a researcher (i.e., its job relevance) can increase when the infrastructures can fully accommodate large-scale data requirements.*

Furthermore, Patel (2016) highlights that data infrastructures should tackle the issues surrounding legal requirements and regulations (such as those of copyright and licensing) as the literature reports these as important barriers to open research data adoption (Zuiderwijk et al., 2020). Regarding copyrights, determining the owner of data may not be easy, not only because the rules on ownership of data could depend on the national laws and regulations, but also because the ownership issue involves many stakeholders ranging from researchers, data

collectors, data analysts, institution or university, and the funding agencies (Patel, 2016). Furthermore, when data sources are copyrighted, this is a barrier to freely sharing datasets (Piwowar et al., 2007). On the other hand, a somewhat deeper issue is regarding licensing: licensing involves restrictions on use, distribution of data, and terms and conditions for derivative works (Patel, 2016). Researchers often see licensing as just a burden and they may experience difficulty in understanding them (Schmidt et al., 2016). *This issue highlights an issue on the perceived ease of use in the TAMs. There is a lack of ease in using open data infrastructures due to heavy burdens of dealing with licensing and copyrighting, and an infrastructural instrument can target to ease this issue.* In that regard, **actively supporting researchers in incorporating licensing and copyright processes** is an important infrastructural instrument. Many data repositories actively engage in license selection processes to help data owners in making the right choices: For example, *4TU.Research*data - an international data repository established by the collaboration between 4 leading technical universities in the Netherlands- explicitly requires data uploaders to select a license as part of the deposit process while clearly explaining different license types (4TU.RESEARCHDATA, n.d.). TU Delft's research data management plan refers users to EUDAT's license selector, which is a tool that helps researchers make a suitable choice of license (*Dmponline-TU Delft*, n.d.).

Another important instrument is integrating different data infrastructures and making them compatible to increase their usability (Strategy for a European Data Infrastructure, n.d.; Zuiderwijk & van Gend, in press). A very prominent example of this is Harvard University's open data initiative Dataverse project's integrated design, which enables many integrations with other systems such as integrations for getting data in (Dropbox, Open journal systems, etc.), for data anonymization (*Amnesia),* and for data analysis (*Data Explorer, Tworavens/Zelig*) (*Harvard Dataverse*, n.d.)**.** *These instruments could support the perceived ease of use of open data infrastructures and therefore could enhance the attitudes towards adoption.*

The second category of instruments relates to making data comply with FAIR principles. FAIR principles are a set of guiding principles to make data Findable (F), Accessible (A), Interoperable (I), and Reusable (R) (Wilkinson et al., 2016). The first principle, making data findable, refers to using a globally unique and eternally persistent identifier, describing data with metadata, placing data in a searchable environment, and specifying the data identifier in the metadata (FORCE11, n.d.; Wilkinson et al., 2016). In line with this, an infrastructural instrument is then **data repositories and other infrastructures facilitating usage of metadata standards** (Shelly & Jackson, 2018; Zuiderwijk, 2020; Zuiderwijk & van Gend, in press). Metadata standards are used to establish a common way of understanding data, and it also prescribes general principles and implementation of such standards (the University of Pittsburgh, n.d.). The difficulty of using standards for data sharing is indicated as a barrier to researchers' willingness toward data sharing practices (Schmidt et al., 2016; Zuiderwijk et al., 2020). Tenopir et al. (2011) argue that many researchers do not use existing metadata standards and instead they opt for creating their own approach to standardization. This results in uneven and inappropriate documentation of data leading to data being indiscoverable or irreproducible

(Michener, 2015; Tenopir et al., 2011). Data repositories should then facilitate the usage of metadata standards; they should enable the storage of metadata on the platform and it should easily let the researcher browse metadata during the data search process (Michener, 2015; Zuiderwijk & van Gend, in press). *This instrument is strongly related to the "result demonstrability" construct (which enhances perceived usefulness) in TAM 2: Researchers are not expected to use an open data infrastructure unless they observe that the infrastructure is beneficial. Furthermore, no benefits can be observed unless the data is findable. Therefore, only when metadata standards are properly adopted in the scientific community, the researcher can see data as discoverable and reproducible on the infrastructure (which means that the infrastructure is presenting interoperable data to the user). Moreover, this instrument also enhances the "output quality" of the system since the quality of the stored data is directly influenced by whether appropriate metadata standards are used.* Michener (2015) argues that establishing data standards is not sufficient for the actual adoption of such standards: **the availability of accompanying software tools supporting metadata creation and management** is also a necessary condition (Michener, 2015; Zhang & Gourley, 2009). An example in the field of atmospheric research is NetCDF (Network Common Data Form), which includes software libraries as well as a data format that enables the creation, sharing, and use of data by application software (Michener, 2015; NetCDF, n.d.).

Moreover, another important instrument is **facilitating researchers to properly create data citations on infrastructures** (Crosas, 2016; Patel, 2016). Citing data is an important driver for data sharing since research shows that getting recognition in the form of increased citations is a motivator for researchers to share their data openly (Zuiderwijk et al., 2020). In *TAM 2, this instrument essentially relates to the "image" construct that influences the perceived usefulness of an innovation. For a potential user of the system (researcher), sharing research data and therefore getting more citations means obtaining "a favorable image" within a reference group (i.e., the research field). Therefore, an instrument that supports this building of an image could be essential in enhancing motivations for using open data infrastructures.* Harvard University's open data initiative Dataverse project stresses implementing proper citation standardization as one of the best practices of data infrastructure (*Harvard Dataverse*, n.d.). For, example when a data owner creates a dataset in the Dataverse repository, the citation is automatically generated on the infrastructure (*Harvard Dataverse*, n.d.). *Facilitating citations could also be valuable in the context of (re)formulating the "subjective norm" towards a higher intention to use the system, as established in the TAMs. If proper citations are done on open data infrastructures, this could lay the groundwork for a new subjective norm in which the belief that data sharing brings rewards is strengthened in research fields.*

Regarding making data findable, an infrastructural instrument is ensuring the use of tools where data from various disciplines can be stored or fetched (*Strategy for a European Data Infrastructure*, n.d.). This concerns the usage of registries like *re3data,* which is a global registry of research data repositories that encompasses research data repositories across different academic disciplines (re3Data.org, n.d.). Zuiderwijk & van Gend (in press) suggest that it is important for **infrastructures to be linked to such "aggregator engines"** to ensure that data are findable regardless of which local repository they are placed under. The overall

integration of different infrastructure elements also implies the need for their overall compatibility and consistency among the data infrastructures. In that regard, European Commission expressed the need to overcome fragmentation issues due to having many data repositories under many different fields: "The landscape of data repositories across Europe is fairly heterogeneous, but there is a solid basis to develop a coherent strategy to overcome the fragmentation and enable research communities to better manage, use, share and preserve data" (European Commission, 2009, p. 7). Connecting infrastructures and data repositories is a requirement to "avoid commuting data" (Behnke et al., 2019). *This instrument is related to the "job relevance" influencing the perceived usefulness of using the system of open data infrastructures in TAM 2. By linking infrastructures to one another, this instrument is in a way ensures that researchers always have easy access to the data which are relevant to them, which would not be the case in case of fragmented and unlinked infrastructures. Therefore, the instrument could reform the technical ecosystem in a way that, regardless of where data is stored, there is always a possibility to find data that are relevant to the researcher, which may change the perceptions of researchers towards comprehending these infrastructures as useful to them.*

Infrastructures should also **explicitly support researchers in the documentation of their data collection methodology** because the better the data collection methodologies are described on repositories, the more willing researchers are to use the openly shared data (Kim & Adler, 2015). These can be done by facilitation of tools such as "open lab notebooks" which help researchers to publish their data as they are creating them, which could increase the quality of methodology documentation (Michener, 2015). *This instrument enhances the "result demonstrability" construct (which enhances perceived usefulness) in TAM 2. Researchers are not expected to use an open data infrastructure unless they observe that the infrastructure is benefiting them. No benefits can be observed unless the data has proper documentation of data collection methodology, because data without an associated data collection methodology is doubtful and not reusable. Therefore, providing support for data collection methodology is expected to positively influence the perceived usefulness of the system according to the TAM 2 model.*

The third category of instruments concerns making the infrastructure secure and trustworthy. Trust is cited as an important aspect to consider in open research data adoption (Zuiderwijk et al., 2020). If a researcher has difficulty in establishing trust in the data that came from someone else, they are less motivated to share and reuse open research data (Zuiderwijk et al., 2020). *As several further augmented versions of TAM (e.g. the models proposed by e.g.* Ghazizadeh et al. (2012) and Ha et al. (2019) *suggest, increasing the trust that is attributed to a system could increase motivations towards using it.* In that regard, **enhancing the trust via the application of certification instruments** could be important in inflicting trust attributed to open data infrastructures such as repositories (Downs, 2021). Data depositors get assurance when they work with a data repository that is assessed and certified as trustworthy (Downs, 2021). In the European Framework for audit and certification of digital repositories, there are three certification instruments, which are *CoreTrustSeal (CTS), Nestor Seal*, and *ISO 16363 certification* (OpenAIRE, 2018). These certification standards are implemented worldwide,

and although they vary in complexity and depth, all these certifications ensure the trustworthiness of a data repository with respect to aspects such as being able to manage the data for future use, data stewardship capabilities, and data curation practices (Downs, 2021; OpenAIRE, 2018). For example, one of the critical aspects that *CoreTrustSeal (CTS)* checks is whether the repository in question can guarantee the integrity and authenticity of the data (OpenAIRE, 2018).

Furthermore, an important obstacle in front of open research data adoption is legal barriers, which are caused by restrictions established by national and international data protection laws concerning processing personal data (Wirth et al., 2021). Two main examples are the US Health Insurance Portability and Accountability Act (HIPAA) and the EU General Data Protection Regulation (GDPR) (Wirth et al., 2021). To comply with such restrictions, researchers often have to anonymize their data before making it openly available on repositories (Childs et al., 2014). Regarding this**, providing researchers with appropriate and user-friendly anonymization tools may be an important element** (Shelly & Jackson, 2018). For example, OpenAIRE provides researchers with a tool called Amnesia, a tool used to flexibly anonymize data according to the GDPR (openAIRE, n.d.). *These instruments also are expected to enhance the perceived ease of use in engaging in open research data sharing and reuse behavior on infrastructures.* Furthermore, infrastructures also have the responsibility **to securely store data, which means that infrastructures and archives need to protect data against hacking, tampering, and unauthorized/accidental deletion** (Patel, 2016). *Similar to tools of certification, these security-related tools also relate to increasing motivation toward using open data infrastructures by enhancing the trust to which users are attributing.*

Table 4 Synthesis of Literature: infrastructural instruments

Instrument Type	Instruments	References
Instruments enhancing the usability of infrastructures	• Infrastructure should be able to accommodate large-scale data, should be easy to use, should enable an increase in data storage growth, and should be reliable	Campbell (2015); Harper & Kim (2018); Kim & Adler (2015); Zuiderwijk et al. (2020)
	• Actively supporting researchers and incorporating licensing and copyright processes	Patel (2016); Piwowar et al. (2007); Schmidt et al. (2016)
	• Infrastructures should be integrated and compatible	Behnke & Staiger (2019; Strategy for a European Data Infrastructure (n.d.); Zuiderwijk & van Gend (in press)
	• Data repositories should be easy to use, and should have user-friendly graphic interfaces	FAIR Data Repositories: Key Features Defined, (n.d.)
	• Data repositories should enable the researcher to do data analysis (as an integrated feature).	da Costa & Lima Leite (2019)

	• Infrastructures should offer assistance in the choice of repository	Downs (2021)
	• Availability of Research Data Management tools (e.g. DMPTool and DMPonline).	Michener, (2015)
Instruments supporting the facilitation of FAIR data principles	• The data repository accommodates and incentivizes the usage of metadata standards: It can store metadata and enable the researcher to browse metadata.	Shelly & Jackson, (2018); Zuiderwijk, (2020); Zuiderwijk & van Gend (in press)
	• Compatibility with different data types and different domain-specific requirements.	Zuiderwijk et al. (2020); Zuiderwijk & van Gend (in press)
	• The data repository inflicts and accommodates proper data citation (standards) so that data can be easily found and attributed.	Crosas, (2016); Patel (2016)
	• Availability of software/tools that are used for metadata creation and management.	Michener (2015); Zhang & Gourley (2009)
	• Adoption of metadata standards	Zuiderwijk et al, (2020)
	• Infrastructures are linked to higher-level search engines/ registry of repositories that enable researchers to search data across different data repositories easily.	FAIR Data Repositories: Key Features Defined (n.d.); Zuiderwijk & van Gend (in press)
	• Providing various query interfaces to accommodate different data search behaviors for the searching functions in infrastructures.	Behnke & Staiger (2019)
	• The open data repository requires the data depositor to provide metadata on the data collection methods. The open data repository stores (meta)data on data collection methods and enables browsing.	Michener (2015)
	• Data usage statistics should be made available on the infrastructures.	Behnke & Staiger (2019)
Instruments concerning security and trust aspects	• Application of certification instruments	Downs (2021)
	• Availability of data anonymization tools	Shelly & Jackson, (2018)
	• Design against accidental data loss	Patel (2016)
	• Infrastructure should be secure against breach	Patel (2016)
	• A variety of access restrictions should be possible on the infrastructure	Behnke & Staiger (2019)

3.2.2.2. Institutional Instruments influencing open research data sharing and reuse

The first category of instruments relates to the governance of research data sharing and reuse processes, which therefore relates to the availability of policies. *This category of instruments interferes with the regulative pillar (i.e. the pillar that encompasses rule-setting activities) in*

the three pillars of institutions framework presented in the institutional theory (Altayar, 2018; Scott, 2013). *By presenting various regulative rules and standardizations ("basis of order"), in the form of policy rules, such instruments create the basis of legitimacy where actors are "sanctioned" when they do not obey them. Regarding policymaking, this happens either in the form of "guilt" or in terms of more formal sanctions such as institutional pressures (coercive mechanisms).* The first instrument is the **availability of data management plans.** A data management plan explains the way the research data are collected, the way the researcher collects the data, and the way the researcher uses the data both during as well as after the research is finalized (Wageningen University & Research, n.d.). In that regard, establishing a plan for managing data causes researchers to consider how they will handle the data and also to think about openly sharing their data (Zuiderwijk et al., 2020). Making data management plans a requirement for the researcher is found to be a motivating factor for a researcher's data sharing behavior (Zuiderwijk et al., 2020). Michener (2015) explains that a data management plan shall cover a variety of topics to ensure concreteness: it should cover data collection and processing methods; organization of collected data; access and use policies; quality control procedures; metadata creation; data preservation; budgeting that explains possible costs regarding preparing, documenting and archiving data; and data sharing plans. The second instrument **is establishing an institutional data sharing policy** (Patel, 2016). A data sharing policy explains rules, principles, and guidelines regarding data governance, data quality, and data architecture throughout an institution (OSTHUS, n.d.). Therefore, the policy should delineate various aspects of data sharing and reuse governance such as purpose and scope of data sharing, guidelines on data submission, guidelines on licensing, metadata entry procedures, data categorizations, possible copyright agreements, conditions about withdrawal requests, terms & conditions of use of data, guidelines on the protection of sensitive data, protection against security breaches; and conditions about intellectual property (Patel, 2016). Such policies are valuable because they help everyone who is involved in research be aware of their rights and responsibilities (Michener, 2015). Furthermore, having a well-defined policy for research data security is valuable because data breaches are common (Patel, 2016). **Having concrete institutional data sharing guidelines** is argued to be an essential instrument for ensuring more collaboration on data-intensive research or implementation of higher-level road maps which are adopted for open science (Ringersma & Adamse, 2019). **Another instrument is explicitly explaining the legal obligations of researchers and explanations of ways for complying with such obligations**. *This instrument differs from the rest of the instruments in this category in the sense that it helps researchers to understand the regulative rules ("basis of order") and possible consequences of (in)compliance ("coercive mechanisms") that already exist in the regulatory pillar, rather than introducing new rules. It can be said that these instruments bring transparency to the regulative pillar so that actors are better informed about their responsibilities.* Since considering licensing aspects are important barriers in front of data sharing behaviors of researchers (Schmidt et al., 2016), it is important to have unambiguous copyright statements and data licenses for use, which could be achieved by ensuring that every research project defines terms and conditions related to the ownership of the data (Patel, 2016). For example, TU Delft's data management plan (*dmponline*) is asking the researcher to establish the choice of licensing and the conditions of data ownership at a very early stage in the research cycle (*Dmponline-TU Delft*, n.d.). For example, on their data management web

page, Radboud University not only explains options for licensing but also refers researchers to the legal departments of the institution or institution's data stewards if they need help. This instrument also relates to explaining strategies for complying with the GDPR.

The second category of instruments refers to ones that actively support researchers during the process of sharing and reusing open research data. Shelly & Jackson (2018) found out that although researchers generally have encouragement to share research data, there is an overall lack of practical support regarding how to engage in such data sharing activities and that there is a growing demand for research data management support from researchers. In this regard, establishing new roles for certain actors or reformulating existing roles should be a goal of several instruments in this category. *(Re)creating roles for certain actors can be considered as an intervention in the normative pillar in the three pillars of institutions framework established in the institutional theory* (Altayar, 2018; Scott, 2013). *These instruments could aim to enhance the obligations of certain actors towards supporting researchers. Therefore, these instruments reintroduce the roles of libraries/ librarians, legal teams, data stewards, data managers, etc., by giving them new goals, activities, and responsibilities, which is the differential basis of order in normative systems* (Scott, 2013). Shelly & Jackson (2018) claim that **libraries should have an active role in engaging with research data management (RDM)** (Shelly & Jackson, 2018). Tenopir et al. (2012) refer to such support from libraries as Research Data Services (RDS). Research shows that currently libraries at universities mostly offer consultation and information services to researchers (Tenopir et al., 2012). Such information and consultation-related support could be, for example, related to reference support for finding, citing research data; and consulting researchers on data management plans or metadata standards (Tenopir et al., 2012). Moreover, they could also be related to **aiding researchers to choose appropriate infrastructures and tools so that they can identify the disciplinary repository most appropriate for their data** (University of Colorado, n.d.)**. Especially the guidance on the selection of repositories should be as early as possible to prevent hasty decisions** later (Downs, 2021). **However, libraries often lack providing technical support to researchers** (Shelly & Jackson, 2018; Tenopir et al., 2012). Such technical support could be in terms of digital curation of data, preparing datasets for a repository, accessing a repository, archiving data, backup practices, removing data from repositories, and creating metadata for datasets (Neylon, 2017; Shelly & Jackson, 2018; Tenopir et al., 2012). Library's supporting role as an institutional instrument should be perceived as versatile, including both technical and non-technical support. Similar to enhancing the supporting roles of libraries, **legal departments can also provide institutional support in terms of privacy, data ownership, and copyright** (Ringersma & Adamse, 2019)**. Ensuring that data stewards whose roles are concretely established are available to the researcher is also an important institutional instrument** (Ringersma & Adamse, 2019). The role of a data steward is also multi-faceted: Ringersma & Adamse (2019) concretely categorize the roles of data stewards into four categories: (1) Policy implementation & compliance (e.g. following the implementation of the Data Management Plans), (2) Services (e.g. advising researcher on data storage environments, data storage standards, file formats, versioning, data documentation, etc. during research), (3) Archiving & Registration (e.g. assisting in metadata creation) and (4) Infrastructure and tools (e.g. advising researcher on infrastructure and tool selection throughout the data/research life cycle). **A**

relatively newer instrument is introducing the data managers, whose main function is to take care of RDM for specific projects (Utrecht University, n.d.; Zuiderwijk & van Gend, in press). In practice, the support from data managers is a paid service, available on a flexible, part-time, or temporary basis (Utrecht University, n.d.). Because data managers are experienced in guidelines and regulations of data sharing, they can save a lot of time for researchers. Introducing data managers means switching research data management tasks from unwilling, inexperienced researchers (or those that do not have time for these tasks) to experienced professionals (Utrecht University, n.d.; Zuiderwijk & van Gend, in press). This is also in line with research findings highlighting that "lack of time" is a fundamental reason why researchers choose not to share data (Tenopir et al., 2018). For example, Utrecht University Library offers a pool of data managers, from which researchers can choose and hire for their research projects at financially attractive prices (Utrecht University, n.d.).

Another institutional instrument is considering moving traditional support services, which exist on more outdated platforms, to more modern platforms such as the webpages (Tenopir et al., 2012). This means considering moving important practical information, which could traditionally only be located in policy documents and long guidelines, to the websites where researchers have higher engagement (Tenopir et al., 2012). Neylon (2017) found out that interventions coming from policy documents have much less influence than those that are more interactive and practical. The author states that providing support may be more valuable than working on details of policy design (Neylon, 2017). Moreover, policy documents are often seen as just another regulatory burden by researchers (Zuiderwijk, 2020). *This finding points out possible problems in creating the basis of legitimacy established in the regulative pillar through policy documents. If policy documents are just seen as a burden, they may have an unwanted, negative effect on motivations toward the behavior (of open data sharing and reuse). If policy documents do not help the researchers or cause researchers to give up on the favored behavior -instead of motivating them-, then, alternative instruments, which are possibly less coercive, should be also established for inflicting change.* Creating web guides for locating data, or placing the requirements of funding agencies' data sharing requirements on institutions' webpages could be such instruments (Tenopir et al., 2012; University of Colorado, n.d.). **It is also useful to place links to full guidelines on the websites** (Shelly & Jackson, 2018).

Furthermore, **another instrument that could support researchers is training and educational support** (Piwowar et al., 2008; Tenopir et al., 2012). *Training and educational support are related to creating normative pressures through the stream of "formal education", which is discussed by* Dimaggio & Powell (1983) *under the context of institutional isomorphism. The instrument of training and educating researchers on various elements of open data sharing aims to create a normative pressure stemming from professionalization, where methods and conditions of a certain work activity (open research data sharing and reuse) are established so that normative and organizational rules about professional behaviors are set and enhanced in an institution* (Dimaggio & Powell, 1983). A case study by Neylon (2017) demonstrated that researchers highly benefited from training in terms of archiving and backup practices. This is also in line with the research findings that indicate researchers' lack

of experience in data sharing practices (Zuiderwijk & Spiers, 2019). The responsibility of training provision could be on various actors, such as the libraries or the specific research groups in universities (Shelly & Jackson, 2018). Nevertheless, it is useful to clarify who is responsible for training services in policy documents to prevent confusion (Shelly & Jackson, 2018). Such training can be provided via seminars given by libraries, creating online data management courses, or referring to existing online courses on tools on institution webpage (Downs, 2021; Shelly & Jackson, 2018; Zuiderwijk, 2020). One specific point where technical training is requested by researchers is training in the digital description and curation of large data sets (Creamer et al., 2012). Piwowar et al. (2008) recommend that data sharing education should be included in the curricula of introductory research courses. Open science curricula could also contain such trainings.

The third category of instruments is those that create financial sources for researchers. An instrument in this regard is **creating appropriate financial resources for data sharing practices, such as providing extra funds for data management** (Piwowar et al., 2008; Zuiderwijk, 2020). When researchers are provided with adequate funding for treatment and management of (open) research data, their motivations for sharing and reusing data increase (da Costa & Lima Leite, 2019; Zuiderwijk et al., 2020). For example, FAIR Data Fund, which is organized by *4TU.Research*data, allows researchers to cover the cost of making data comply with the FAIR principles (e.g. by implementing the appropriate metadata standards or properly anonymizing the datasets) (4TU.ResearchData, n.d.). Zuiderwijk (2020) highlights that there could be a visibility problem concerning such funds, meaning that researchers may not be aware of such funds' availability. Therefore, promoting the existence of such funds could also be an important instrument. Regardless, covering the costs of data sharing remains valuable (Piwowar et al., 2008).

The final category of instruments relates to building a data sharing culture and creating incentives. *This category of instruments interferes with the cultural-cognitive pillar in the three pillars of institutions framework* (Altayar, 2018; Scott, 2013). *According to the institutional theory, building a culture of data sharing can be possible by building a shared understanding in the social group as a basis of compliance* (Scott, 2013). *In the cultural-cognitive pillar, it is believed that people act the way they act because they are not aware of other ways of acting* (Scott, 2013). *Therefore, researchers may not be aware of an alternative research culture where data sharing is a fundamental goal*, *which means that instruments should target reestablishing goals.* In that regard, one specific way to achieve a shared understanding of a data sharing culture is by **revising policies and guidelines to reflect data sharing goals** so that institutional policies are explicitly incentivizing data sharing (Patel, 2016; Piwowar et al., 2008).

Furthermore, creating a data sharing culture could also be possible by placing enough incentives in the form of formal or informal rules (Scott, 2013). *A way to create motivations and influence behavior toward data sharing and reuse could be via introducing appropriate rewards as established in the regulatory pillar* (Scott, 2013). **Recognizing and rewarding data sharing contributions is crucial**, **since the lack of recognition is cited as an**

(unresolved) issue in the literature (Piwowar et al., 2008). **Using track metrics for data sharing contributions** in academic research could be a powerful tool (Piwowar et al., 2008; Zuiderwijk & van Gend, in press). Furthermore, Piwowar et al. (2008) **recommend that in research institutions, data sharing contributions should be considered during hiring, tenure, and promotion decisions; which could be possible, for instance, by providing a bonus to a research paper's impact factor if the associated research data are made openly available**. **Adoption of data citation policies** that ensure research data are cited just like other types of publications could also indirectly result in more rewards. (WILEY, n.d.).

Other instruments in this category refer to introducing requests for data sharing. *Requests could be perceived as a creation of coercive mechanisms influencing behavior towards open research data adoption in the regulatory pillar.* A relevant instrument is **incentivizing data sharing practices by publication policies**. This can be done by scientific journals making data sharing a mandatory requirement (Michener, 2015; Patel, 2016; Piwowar et al., 2008). For example, when journals in the field of Evolution and Ecology adopted the Joint Data Archiving Policy (JDAP), which is a policy that requires authors to share their research data to support their findings, a significant increase in data sharing in these fields was observed (Michener, 2015). **A similar instrument is incentivizing data sharing practices by funders, again, in the form of a (mandatory) requirement** (Michener, 2015; Patel, 2016; Piwowar et al., 2008). Funders can also incentivize research data sharing by evaluating a proposal's data sharing plan under its scientific contribution (Piwowar et al., 2008).

Another instrument that facilitates data sharing culture is **demonstrating the benefits of and needs for data sharing, and also how the issues around privacy and data ownership can be tackled** (Piwowar et al., 2008). Especially the issue of data ownership is heavily cited in the literature as one of the barriers in front of open research data adoption (Zuiderwijk et al., 2020). Therefore, the concept of data ownership should be properly described to researchers. Piwowar et al. (2008) recommend that universities may hold seminars where researchers can be prompted to think about how to maintain privacy while maximizing scientific benefit in data sharing; and to change their mindset from "data ownership" to "data control". *According to the institutional theory, such seminars could be appropriate platforms for the presentation of rationalized myths of open research data sharing as an innovation. By communicating the benefits of open research data sharing and reuse to the researchers, policymakers could* exert *institutional pressure and lead to organization-wide changes* (Altayar, 2018; Meyer & Rowan, 1977). *Although benefits may not be tangible in the eyes of the researcher at the beginning of adoption, appropriate rationalization can influence behavior and therefore lead to the actual realization of the benefits.* Similarly, another instrument that allows for communicating rationalized myths is research organizations **actively publishing experiences in data sharing to incentivize researchers,** as showcasing the benefits of data sharing to the researcher in close contact could be a strong driver for adoption (Piwowar et al., 2008). An institution's website can be a good platform to acknowledge data sharing and reuse efforts (Zuiderwijk & van Gend, in press).

Table 5 Synthesis of Literature: institutional instruments

Instrument Category	Instrument	Reference
Instruments that manage and govern data sharing and use process	• Establishing concrete data management plans	Michener (2015); Shelly & Jackson, (2018); Tenopir et al. (2012)
	• Establishing institutional data sharing policy and guidelines for data sharing	Michener (2015); Patel (2016); Shelly & Jackson (2018)
	• Establishing policies for research data security	Patel (2016)
	• Ensuring that researchers think about costs related to access, management, and preservation of data before the research starts.	OECD (2007)
	• Establishing a data deletion policy.	Behnke & Staiger (2019); FAIR Data Repositories: Key Features Defined (n.d.)
	• Giving researchers a clear legal basis about rights of use, so that they understand what they are allowed to do with the data; explaining legal requirements and options of compliance to such requirements; asking the researcher to clarify terms of use at the beginning of the research cycle (e.g. concerning licensing, privacy confidentiality).	Dmponline-TU Delft, (n.d.); Fecher et al., (2015); Patel (2016); Zuiderwijk & van Gend (in press)
	• Supporting the alignment of organizational data sharing and management policies between organizations and countries.	(Clarke & Davidson, 2021)
	• Giving a clear guideline on how to obtain consent for data sharing.	Fecher et al. 2015)
	• Giving a clear guideline on how to anonymize data.	Fecher et al. (2015)
Instruments that actively support researchers in sharing and using research data	• Establishing support from libraries, clarify the role of libraries.	Neylon (2017); Shelly & Jackson (2018); Tenopir et al. (2012); Zuiderwijk (2020)
	• Providing guidance on the selection of data repository as early as possible in the research cycle.	Downs (2021); Zuiderwijk & van Gend (in press)
	• Establishing support from the legal teams of the organization.	Ringersma & Adamse (2019)
	• Placing practical information (e.g. guides on locating data or funding agency requirements) on webpages and web guides; carrying information from traditional information platforms (e.g. documents) to web pages.	Neylon (2017); Shelly & Jackson (2018); Tenopir et al. (2012); Zuiderwijk (2020)
	• Educational support on data management: providing training to researchers	Neylon (2017); Piwowar et al. (2008); Shelly & Jackson (2018); Tenopir et al. (2012); Zuiderwijk (2020)
	• Availability of data stewards whose roles are concretely established in the organization.	Ringersma & Adamse (2019)
	• Possibility of working with data managers to shift responsibility from researcher to an experienced professional.	Utrecht University (n.d.); Zuiderwijk & van Gend (in press)
Instruments that relate to financial resources	• Providing financial support to researchers and make the availability of funding clear.	da Costa & Lima Leite (2019); Piwowar et al. (2008); Zuiderwijk (2020)
Instruments that build a culture of data	• Revising policies and guidelines in an institution to reflect data sharing goals	Patel (2016); Piwowar et al. (2008)

sharing and create incentives	• Recognizing and rewarding data sharing contributions (e.g. via track metrics)	Piwowar et al. (2008); Zuiderwijk & van Gend (in press)
	• Data sharing contributions should be considered during hiring, tenure, and promotion decisions.	(Piwowar et al., 2008)
	• Implementing data citation policies	WILEY (n.d.)
	• Demonstration of benefits of and needs for data sharing	Piwowar et al. (2008)
	• Demonstration of how the issues around data ownership and privacy can be tackled.	Piwowar et al. (2008)
	• Creating incentivizes from publishers or from organizations (e.g. requests for sharing data)	Michener (2015); Piwowar et al. (2008); Zuiderwijk et al. (2020)
	• Creating incentivizes from funders (e.g. requests for sharing data or by evaluating a proposal's data sharing plan under its scientific contribution)	Michener (2015); Patel (2016); Piwowar et al. (2008)
	• Actively publishing experiences in data sharing to incentivize researchers	Piwowar et al. (2008); Zuiderwijk & van Gend (in press)

3.3. Conceptual Framework

From the analysis presented in the previous section, the following conceptual frameworks (Figure 8 and Figure 9) are formulated, and they are taken as basis for the case study.

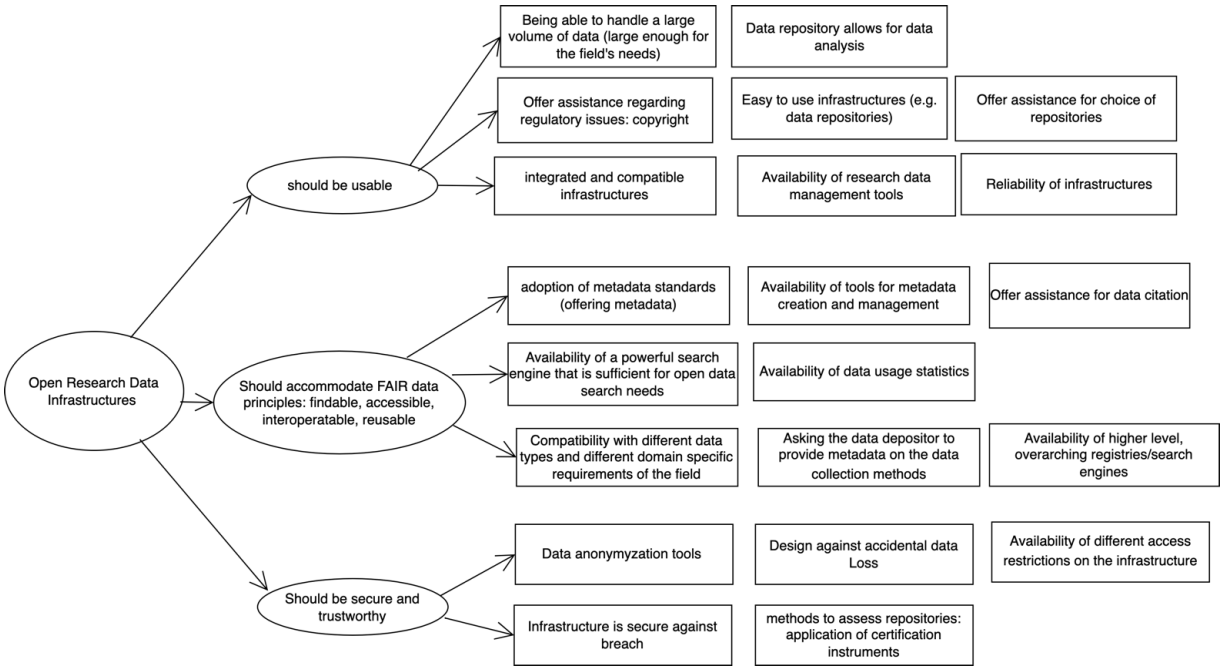

Figure 8 Conceptualization of infrastructural instruments influencing open research data sharing and reuse

Figure 9 Conceptualization of institutional instruments influencing open research data sharing and reuse

4. The case study

This chapter explains the case study that is conducted to understand the influence of the proposed infrastructural and institutional instruments on open research data sharing and reuse practices in the field of Epidemiology. The chapter first describes the motivation for choosing a case study approach, the case study selection criteria, the case study information sources, and the design of the interviews. Furthermore, it explains how the qualitative data are analyzed and operationalized in this study to ensure a scientific approach. It then describes the background of the case study, and finally presents the findings of the case study.

4.1. Motivation for case study approach

Yin (2018) defines the case study research method as "an empirical method that investigates a contemporary phenomenon (the "case") in depth and within its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident" (Yin, 2018, p. 45). He then suggests following a case study approach when three conditions are met: (1) when the main research questions are "how" or "why" questions, (2) there is not much control over the behavioral events that form the system that is questioned, and (3) the focus of the research is contemporary rather than entirely historical events (Yin, 2018).

As previously described in chapter 2, the case study aims to answer the specific "how" type of (sub)question, which is "*how do infrastructural and institutional instruments influence researchers in openly sharing their research data and in using openly available research data in the field of Epidemiology?*". Furthermore, the research nature is highly exploratory because it is still unclear which instruments have actual influence in the field. The research is not interested in mere frequencies of incidents but rather an analysis of "why" a certain process (open data practice) occurs the way it occurs (Yin, 2018). Moreover, considering the researcher's position relative to the system, there is no control over the relevant behaviors (of the Epidemiology researchers), meaning that the author of this study does not influence the behaviors emerging in the system (Yin, 2018; Zuiderwijk & van Gend, in press). This confirms the study's case study approach. Finally, if one views open research data practices as a novel phenomenon, then the focus of this research is not solely on the present, but rather on the rendition of the present and the recent past, meaning that the study analyzes open data practices while considering its novelty (i.e. its difference from the recent past as well as its relation to the recent past) (Yin, 2018). This makes the focus of the research a contemporary process, which also confirms the case study approach (Yin, 2018).

4.2. Case Study Selection

In the case study research approach, apart from building appropriate questions, an important component is identifying the case to be studied (Yin, 2018). To accomplish this, identifying the case and defining its boundary are important steps (Yin, 2018). In this study, a single case study approach is adopted because the study is interested in understanding open data practices in a specific "community" (i.e. the field of Epidemiology) (Mohd Ishak & Abu Bakar, 2014).

The following are the case study selection criteria that are used to select the case:

1. The case focuses on open research data sharing and reuse (mainly) in the Netherlands. As the author of this study is a student of a Dutch University (TU Delft), focusing on a Dutch case is ideal for this research. This criterion also means that the case focuses on open data in the context of "research", implying that the interviewees should be engaging with research.
2. The case focuses on a field where the open research data practices are expected to be at low levels. As the previous sections of this report already suggested and explained, the qualifying field for this is the field of Epidemiology.
3. The case should allow for finding information sources (e.g. interviewees) from different organizations under the defined geographical boundary (i.e. the Netherlands) and the field (Epidemiology). This is important because this study aims to discuss ways to enhance open data practices in a certain field, rather than in a certain organization. Therefore, being able to find interviewees across different organizations is an important criterion for case selection.
4. The case enables access to interviewees with at least low levels of previous experience in open data practices. This is important because having no experience in open data practices would eliminate the chance to discuss infrastructural and institutional instruments with interviewees. (During the study, one researcher reported having no open research data experience. Although not fully in line with the case study selection criteria, this participant was still included in the case because the associated interview provided valuable insights on barriers to open data practices in the field).

Based on the criteria above, the selected is a case study of open research data sharing and reuse in the field of Epidemiology.

4.3. Case study information sources

The main information source of this case is interviews with researchers working in the field of Epidemiology and a research data management consultant[2]. This is complemented with analyses of websites of the organizations where these researchers work and their policy documents.

It is important to have participants across different organizations for the case study to have external validity (i.e. that the case study's findings can be generalized to the entire field). For

[2] During the interviews with Epidemiology researchers, it was brought up by several researchers that the legal context forming the boundary of data sharing in Epidemiology influences open data practices heavily. It was mentioned several times that open research data for the field of human health is usually not fully guided by individual researchers' behavior, but it is also significantly bounded by GDPR privacy laws and informed consent procedures. Therefore, in light of these developments during the interviews, a research data management consultant who has expertise in the legal/privacy aspects of open data practices and research data management was also added as a case study information source.

this reason, the quota sampling approach is used to recruit Epidemiology researchers for the study (Mohd Ishak & Abu Bakar, 2014). This approach is suitable when there is a need to interview a group of people with different characteristics to make sure that there are some specific differences in the sample (Mohd Ishak & Abu Bakar, 2014). This is because the unit of analysis in this study is individuals (researchers) of different characteristics (working for different organizations), clustered under one group (the field of Epidemiology) (Mohd Ishak & Abu Bakar, 2014). In line with this, participants were recruited across different research institutions with distinct Epidemiology departments in the Netherlands. Possible participants were initially detected by using the websites of different research institutions (e.g. university medical centers). Furthermore, some other participants were detected through the personal network of one of the supervisors of this study. To recruit the participants, a draft email was written, which can be seen in Appendix A. Initially, around twenty emails were sent. However, the initial response rate was very low, around three people responded at first and only one of these people expressed a willingness to participate in the study. This is why new batches of emails were sent till a sufficient number of participants (minimum of ten) for the study was reached.

Table 6 Case Study Information Sources

Case study information sources	**References**
Interviews with Epidemiology researchers	Interviews with ten Epidemiology researchers
Interview with a research data management consultant	An interview with a research data management consultant who has expertise in legal aspects of open research data practices and research data management
Policy documents	UMC Utrecht Research Data Management Policy[3], Leiden University Data Management Regulations[4]
Websites	UMC Utrecht Research Data Sharing webpage[5], Amsterdam UMC Research Data Management support webpage[6], Utrecht University Research Data Management Support Webpage[7], Utrecht University Recognition and Rewards Webpage[8]

Ten Epidemiology researchers with varying positions, subfields of Epidemiology, and ages were interviewed. The background information on these participants can be seen in Table 7. The male/female division of the interviewees is 6/5. The participants work in the following institutions: Leiden University Medical Centre (LUMC), University Medical Centre Groningen (UMCG), Amsterdam University Medical Centre, Utrecht University, and University Medical Centre Utrecht. Out of ten Epidemiology researchers interviewed, nine of them reported having

[3] https://www.uu.nl/sites/default/files/rdmpolicy_umcu_eng_v3.1.pdf
[4]https://www.organisatiegids.universiteitleiden.nl/binaries/content/assets/ul2staff/reglementen/onderzoek/research-data-management-regulations-leiden-university.pdf
[5] https://www.umcutrecht.nl/en/research-data-umc-utrecht
[6] https://www.amsterdamumc.org/en/research/support/about/research-data-management.htm
[7] https://www.uu.nl/en/research/research-data-management/guides/policies-codes-of-conduct-and-laws#ownership
[8] https://www.uu.nl/en/research/open-science/tracks/recognition-and-rewards

previous experiences with open research data practices (openly sharing research data and/or reusing openly shared research data).

Table 7 Background information on the interviewees

Interviewee no.	Role of interviewee	Age group	Academic Level	Experience with open research data sharing and/or reuse
I1	Epidemiology Researcher	25-30	Ph.D. student	Experience with open research data reuse
I2	Epidemiology Researcher (and Policy Advisor)	36-40	Assistant Professor	Experience with open research data sharing and reuse
I3	Epidemiology Researcher	56-60	Full professor	Experience with open research data reuse
I4	Epidemiology Researcher	31-35	Postdoctoral Researcher	Experience with open research data reuse
I5	Epidemiology Researcher	25-30	Postdoctoral Researcher	Experience with open research data sharing and reuse
I6	Epidemiology Researcher	25-30	Ph.D. student	No experience in open research data practices.
I7	Epidemiology Researcher	25-30	Postdoctoral Researcher	Experience with open research data reuse
I8	Epidemiology Researcher	41-45	Assistant Professor	Experience with open research data sharing and reuse.
I9	Epidemiology Researcher	41-45	Associate Professor	Experience with open research data sharing and reuse
I10	Epidemiology Researcher	46-50	Associate Professor	Experience with open research data sharing
I11	Research Data Management Consultant			Expertise in the legal/privacy aspects of open data practices and research data management (in life sciences and also in other research fields)

4.4. Interview design

Interview questions were formulated using the conceptual framework that was illustrated in section 3.3. The interview was divided into five distinct sections: (1) Background information, (2) Previous experiences in open research data sharing and reuse, (3) Infrastructural instruments that influence motivation and behavior towards open data practices, (4) Institutional instruments that influence motivation and behavior towards open data practices, and (5) Barriers to open research data sharing and reuse. The full interview questions can be seen in Appendix B.

Regarding infrastructural and institutional instruments, section three and section four provided statements that each contained an instrument, and asked the respondent to explain to what extent they have access to such an instrument and how such an instrument may affect open data practices. The purpose was to understand (1) whether the instrument was available to the researcher and (2) whether this instrument is an important factor for open data practices. To get participants familiar with the context of the research and with these instruments, brief definitions of open data infrastructures as well as institutions were provided to the participants.

The interview was tested separately with two different TU Delft students (one master's and one bachelor's student) who also are currently studying the topic of open data for their master's and bachelor's theses. Testing the interviews was useful because it made it clear that various changes can be made to the text of the interview, especially to increase understanding and avoid confusion. For example, upon the advice from the interview testers, the statements under section three and section four were re-ordered to create a logical flow of instrument topics. Moreover, the question *"To what extent do you (dis)agree?"* in sections three and four were changed to *"Do you agree with the statement? (yes, no, partially) Please explain why."* to increase clarity. This change proved to be beneficial later in the research process when we operationalized the data (see chapter 4.5.4.). Furthermore, it was also pointed out that questions that were asking participants to come up with *new* instruments were too demanding. Thus, these questions were revised so that participants are instead asked to point out what kind of features they wish open data infrastructures had, what kind of support they wish their organization provided, or alternatively, what the troublesome issues about the open data infrastructures or their organization are.

For the interview with the research data management consultant, a different interview document was prepared. This interview had a separate section on GDPR and other regulations that affect (open) data sharing epidemiology. These interview questions can be seen in Appendix C.

The interviews with Epidemiology researchers were conducted from March 28 until April 19, 2022. The interview with the research data manager was conducted on May 3, 2022. The interview lengths varied from 30 minutes to 90 minutes. The interviews were conducted in the online setting, using TEAMS software. The data was collected based on informed consent. In line with the informed consent protocol (which all participants signed), all participants agreed to be recorded. The interviews were recorded using the built-in TEAMS recording feature. The built-in transcription feature of TEAMS software automatically generated transcripts of the interviews.

4.5. Analysis of the interviews

4.5.1. Coding of the interviews

From the transcriptions obtained from the TEAMS transcription feature, extensive notes of the interviews were written to analyze the interviews. When writing these notes, no changes were made to the order of the topics discussed in the interview, thus the notes also contained the same sections that the interviews had. For each of the instruments and barriers discussed, the participant's own sentences are summarized in these notes. We gave participants a chance to review these notes, and the participants approved the notes. The interview notes were used in the coding process. Upon reasonable request, the anonymized interview notes are available from the corresponding author, B.O. Türk. The codebook underlying this study is openly available in 4TU.ResearchData repository at http://doi.org/10.4121/20085560

Coding is a central element of qualitative analysis, and it involves examining a portion of the qualitative data by labeling it with a phrase or word that aims to summarize that piece of information (Linneberg & Korsgaard, 2019). The advantages of coding include acquiring deep insights into the data, making the data retrievable, structuring the data, ensuring transparency, ensuring validity, and understanding participants' views from their own lens (perspectives) (Linneberg & Korsgaard, 2019).

For coding the qualitative data in this report, we used the qualitative coding guideline from Linneberg & Korsgaard (2019). According to this guideline, the first step is preparing for the coding process, which includes examining the research question and the research objectives, doing a literature review on the research topic, and preparing the documents on which the coding analysis will be done (Linneberg & Korsgaard, 2019). The second step of the guideline is deciding on which coding approach will be taken (i.e. inductive or deductive). Linneberg & Korsgaard define the inductive coding approach as "codes from the data by using phrases or terms used by the participants themselves, rather than using the, often theoretical, vocabulary of the researcher." (2019, p. 12). This refers to the task of building theory from the data when theoretical concepts are not available to comprehend the phenomenon that the researcher is studying (Linneberg & Korsgaard, 2019). It involves breaking the data and then abstracting it to a higher level, which is "theory building" (Gehman et al., 2018; Linneberg & Korsgaard, 2019). Deductive coding on the other hand refers to creating a pre-defined list of codes in a coding frame before the coding process starts (Miles et al., 2013). This approach is useful when the researcher is testing a theory in qualitative research. Linneberg & Korsgaard state "if the study is theory-driven, the theoretical framework may be converted into a coding framework" (2019, p. 13). Therefore, the codes in deductive coding could be theoretical concepts or themes that are taken from the literature (Linneberg & Korsgaard, 2019).

The next step of the coding process is executing the coding cycles, which often consist of two or more cycles depending on how extensive the research is (Linneberg & Korsgaard, 2019). For each coding cycle, the researcher should choose the code types (e.g. descriptive versus attribute codes) (Linneberg & Korsgaard, 2019). Descriptive codes are codes that are used to explain the parts of the data based on what that part of the data is about, which refers to "using a label that indicates the meaning of the segment of data in relation to the overall research topic" (Linneberg & Korsgaard, 2019). Attribute codes are on the other hand basic information that is assigned for certain parts of the qualitative data (Linneberg & Korsgaard, 2019). For example, this could refer to assigning codes for aspects such as age, gender, experience, or other attributes which the researchers find valuable for the analysis of the study (Linneberg & Korsgaard, 2019). The first cycle of coding, otherwise known as open coding, refers to the examination of the qualitative data, and assigning the segments of the data to codes that capture their essence (*Qualitative Data Analysis: The Research Guide*, 2020). The second cycle focuses on identifying patterns, rules, or cause-effect progressions, which help the researcher understand which parts of the data should be put aside, and which parts should be reexamined to arrive at emerging codes (*Qualitative Data Analysis: The Research Guide*, 2020). Through having these distinct cycles, there is a possibility for making an initial analysis (first cycle), and then looking for overarching structures which can be observed at theoretical levels via

processes of integration, prioritization, abstraction, or conceptualization (second cycle) (Linneberg & Korsgaard, 2019). Therefore, there is a possibility to start from data and end up with a theory by clearly establishing the grounds on which the researcher arrives at conclusions (Gioia et al., 2013; Linneberg & Korsgaard, 2019). For the coding process of our case study, we used the steps that we explain in this subsection. A visual illustration of this process can be seen in Figure 10.

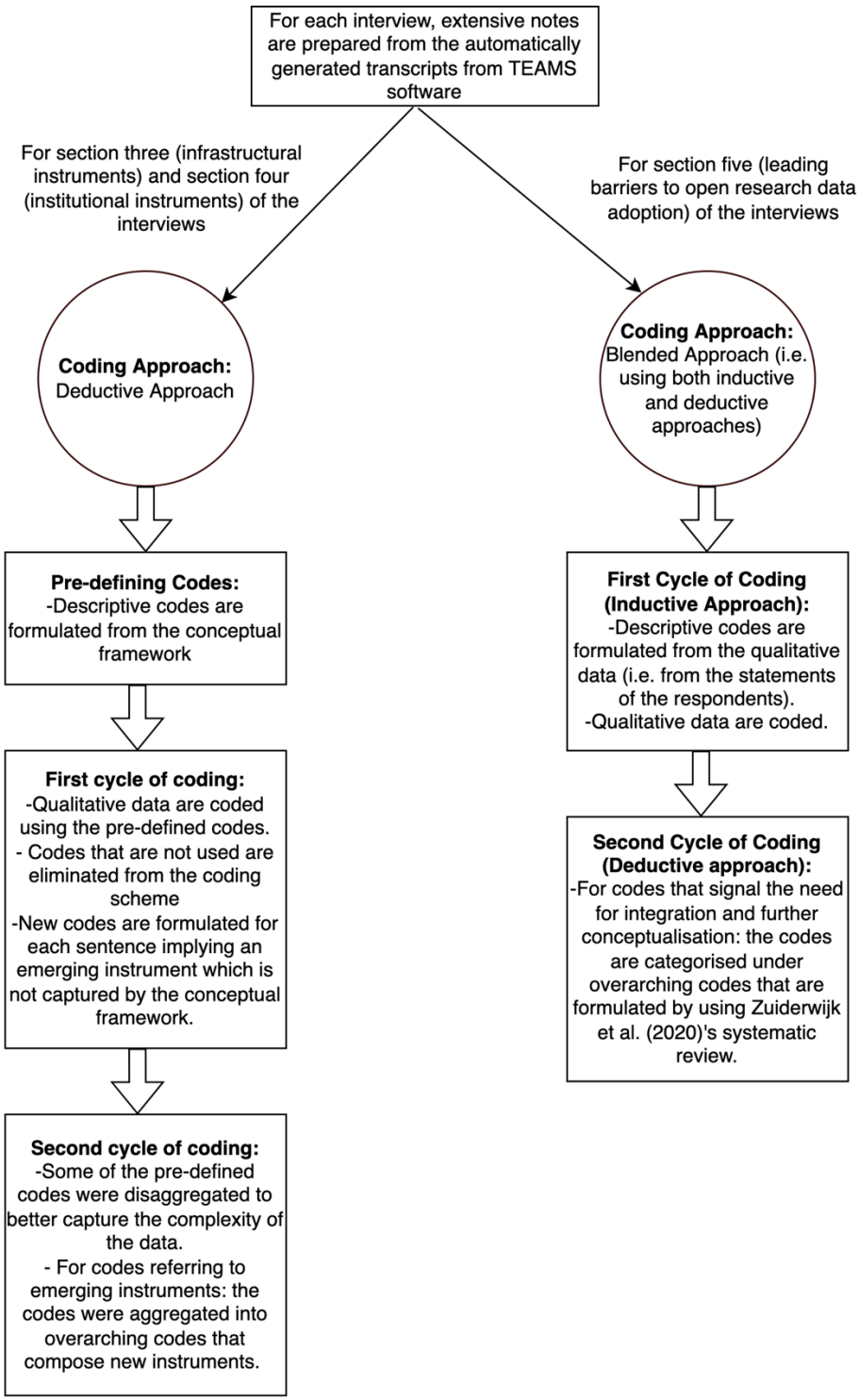

Figure 10 Coding steps that are executed in this study

4.5.2. Coding the qualitative data on infrastructural and institutional instruments

To prepare for the coding process, we reviewed the research question, the interview design, and the conceptual framework established in chapter 3.3. which forms the basis of the interviews. We chose to apply deductive coding when analyzing the second and third sections of the interviews, which are on infrastructural and institutional instruments. Considering that this project is a master's thesis project that has a limited timeframe for completion, the deductive coding approach is favored for the main part of the interviews since the approach prevents the whole process from becoming too complicated and lacking in focus, compared to the inductive coding approach (Linneberg & Korsgaard, 2019).

Using this approach, we prepared a pre-defined list of codes before starting to code the data. Since we had prepared the interview questions (in section three and section four) strictly based on the instruments that are established in the conceptual frameworks in chapter 3.3., we prepared the descriptive codes based on the frameworks. This approach enabled us to place a huge chunk of the qualitative data in the codes that summarized each of the instruments. Some of the instruments were not discussed in the interviews. There are two interrelated reasons for this: (1) Some interviewees did not have any ideas about these instruments and chose to skip them and (2) In some interviews (i.e. particularly the interviews that we conducted the last), there was a lack of time since these interviewees said they could spare only half an hour for the interview, so we excluded the instruments that had not been discussed to that point. We deleted the codes for these instruments from the initial round of coding. For each of the other sentences that implied emerging instruments (that did not exist in our pre-defined list of codes), we created a separate code.

In the second cycle of coding, we realized that some of the codes that referred to an instrument were too broad, and needed to be further disaggregated to discover the diversity of the data and properly conceptualize them (Linneberg & Korsgaard, 2019). Therefore, we further split these codes, such as the code "metadata" which was split into "metadata access" and "metadata on data collection", realizing that the statements given by the participants differed in whether they were talking about the existence of metadata or the content of metadata. Furthermore, in the second cycle of coding, we tried to find the patterns and overarching structures in the chunks of data labeled under the codes that had emerged from the data (which would imply new instruments) as part of the guideline that we described in the previous subsection. Since these codes were rather too specific, we tried to relabel codes into overarching codes in order to progress from data to theory so that we would arrive at new instruments in this process.

4.5.3. Coding the qualitative data on the leading barriers to open data practices

For coding the last section (section five) of the interviews, which is about the leading barriers to open data practices, we followed a combination of inductive and deductive coding approaches, which is called a "blended" or "abduction" approach (Linneberg & Korsgaard, 2019). Using both approaches enables the researcher to go back and forth between data and theory (Linneberg & Korsgaard, 2019). Although there is extensive literature on barriers to

open research data, there is no literature specifically on the barriers for this particular research discipline (Epidemiology). Acknowledging this, we started the first cycle of coding with an inductive approach and developed codes using phrases or words used by the respondents as we did not want to steer the statements of the interviewees towards the general literature. This approach allowed us to stay closer to data compared to the inductive coding approach (Linneberg & Korsgaard, 2019). Creating codes focusing on the respondents' wording resulted in a long list of codes. However, we realized that some of the codes are highly related to another and signaled the need for integration and further conceptualization. However, integrating and further conceptualizing these codes were not possible without using any theory. Linneberg & Korsgaard (2019) state that in the abduction approach, by going from the inductive to deductive approach, a researcher can get closer to the theory. As we aimed to move the coding process toward higher-level categories (which we will later introduce as the *leading barriers* in our case study description), we chose to use the deductive coding approach in the second cycle of coding. Therefore, we integrated the codes that are highly related to one another under labels that are already established by the literature. We used the extensive systematic review of Zuiderwijk et al. (2020) on the factors influencing open research data adoption which helped us conceptualize the barriers we detected in the study. The final set of codes that are reached upon the completion of the coding processes can be seen in the codebook.

4.5.4. From coding to analysis: operationalization

Rao & Reddy (2013) argue that conceptualization and operationalization are the two important processes that researchers who conduct empirical social research use to establish the linkage between concepts and data. Concepts are abstract ideas that represent the (social) phenomenon that the researcher is studying as part of the research (Rao & Reddy, 2013). Since concepts may have different meanings, conceptualization can be defined as "the process through which the researcher attempts to arrive at a common agreement on the meaning of the concepts under study" (Rao & Reddy, 2013, p.109). Rao & Reddy (2013) state that in the research process, conceptualization processes could refer to coming up with a research problem, defining the concepts, reviewing the literature, and building the nominal definition of the meanings of the concepts. In this study, so far, we already executed the conceptualization process. For example, we provided concrete definitions for our research context (e.g., our definition of open research data sharing) in chapter 1. Furthermore, we clarified the research problem and boundary in chapter 2, and we also defined institutional and infrastructural instruments means in the context of open research data sharing and reuse in chapter 3 via our systematic literature review. Therefore, we established definitions of concepts to identify the focus of the study and describe the social phenomena (Rao & Reddy, 2013). We also communicated these nominal definitions to the interviewees by giving brief explanations during the interviews (see Appendix B).

What concerns this subsection of the report is the second vital process in empirical research: operationalization, which is when the researcher creates an operational definition for the concept as well as the steps and procedures that can be used in measuring the concept (Rao & Reddy, 2013). This concerns moving from abstract levels to empirical reality by exchanging concepts with variables and by formulating specific research procedures (Rao & Reddy, 2013).

These procedures will lead to empirical observations and also to the interpretation of the data to answer the research problems of the case study (Rao & Reddy, 2013).

In this case study, our research (sub)question is "*how do infrastructural and institutional instruments influence researchers in openly sharing their research data and in using openly available research data in the field of Epidemiology?*". We propose that the answer to this question of "influence" lies in interpreting (1) whether these instruments are available to the researchers **[availability],** and (2) the extent to which these instruments influence open research data sharing and reuse (i.e. open data practices) **[importance].**

Therefore, in this subsection, we aim to explain how we operationalized "the availability to instruments", their "importance on open data practices" and finally the "influence of instruments on open data practices". In other words, we explain how we converted the dimensions of these concepts into directly measurable entities to generalize our findings to the defined population in our case study (Rao & Reddy, 2013).

4.5.4.1. Availability of an instrument

To measure the availability of each instrument, we examined the answers given in section two and section three of the interviews where we had read a statement to the respondents (for each of the instruments) and asked whether they agree with it in a binary (yes/no) manner ("Do you agree with the statement?"). These statements described being able to use or apply the instrument in question in practice. For example, for the instrument "*Availability of a search engine that is sufficient for open data search needs*", the statement was "*The search engine on the open data repository that I use is sufficient for my open data search needs*". Getting data in this manner enabled us to operationalize the availability of the instruments easily. We then classified the respective statements of respondents under "Yes" or "No", and finally put this information in the codebook. The respondents who could not give a definitive answer to this question (e.g. because they were not sure) were omitted from this process to ensure validity.

4.5.4.2. Importance of an instrument

Measuring the importance of the instruments on open data practices requires more effort, as "importance" itself is rather an abstract concept that cannot directly be asked. Since importance is an abstract concept, we tried to measure it by performing a systematic examination of the answers given by the respondents to the question "*To what extent does this instrument influence your open research data sharing and reuse behavior?"* which we had asked right after the question about the availability for each of the instruments. Therefore, by examining each statement, we aimed at understanding whether the respondent finds the instrument important for open data practices. To achieve this, we classified the respective statements of respondents under "**[the instrument] is an important factor for open data practices"** and noted the number of participants in the codebook when the statement of the respondent fell under one or more of the conditions below:

1. The respondent explicitly mentions that the instrument has an "influence" on open data practices (negative or positive), or that the instrument is "important" or "valuable" or "useful" for open data practices.
2. The respondent explains a concrete causal relationship of how the instrument influences open data practices.
3. The respondent states that there is a "need" for such an instrument for open data practices or better open data practices.
4. The respondent states that they would like to have access to this instrument for their open data practices or that they are "happy" or "satisfied" by already having access to it.
5. The respondent states that if this instrument existed, the level of open data practices would be affected.

We classified the statements of respondents to **"[the instrument] is <u>not</u> an important factor for open data practices**" and noted the number of participants in the codebook when the statement of the respondent fell under one or more of the conditions below;

1. The respondent explicitly mentions that the instrument does not affect open data practices at all, or it affects at low levels.
2. The respondent explains that there is not a relationship between (the existence of) the instrument and open data practices, or that the relationship is highly doubtful or highly questionable.
3. The respondent states that not having access to this instrument is not a (strong) barrier to open data practices.
4. The respondent states that researchers do not need this instrument for their open data practices.
5. The respondent states that researchers are not interested in (using) this instrument regarding open data practices, or that they choose not to use or engage with the instrument even if they have access to it (or would have access to it).

4.5.4.3. Influence of the instrument on open data practices in our case

Finally, we evaluated whether the instrument in question has an influence on open research data practices in our case by using the following strategy**:**

For each instrument, we looked at the <u>difference</u> between the number of respondents stating that having the instrument **is** an important factor for open data practices and the number of respondents stating that the instrument **is not** an important factor for open data practices

[# of respondents stating that having the instrument is an important factor for open data practices - # of respondents stating that having the instrument is not an important factor for open data practices] = X

- If this (X) number is equal to or smaller than 1, we classified the instrument as one that has a low influence on open research data practices in our case.

- If this (X) number is 2, we classified the instrument as one that has a medium influence on open research data practices in our case.
- If this (X) number is between 2 and 5, we classified the instrument as one that has a high influence on open research data practices in our case.
- If this (X) number is equal to or larger than 5, we classified the instrument as one that has a very high influence on open research data practices in our case.

It is important to mention that, although this research is qualitative, this operationalization is rather a quantitative approach. We recognize that this analysis does not imply anything about how important an instrument is in a larger population.

Below is an example from the codebook for the instrument "Availability of a search engine that is sufficient for open data search needs". The full work can be found in the codebook.

Table 8 Excerpt from codebook that explains the operationalization process that feeds the analysis of the case study

Instrument	**Explanation**	**Example of answers given**	**Availability of the instrument**	**Importance of the instrument**	**Influence of the instrument on open research data practices in the case**
Availability of a search engine that is sufficient for open data search needs	Researchers were asked: (1) whether the search engine on the open data repository that they use is sufficient for their open data search needs, (2) the extent to which having a search engine that performs sufficiently influences open research data sharing and reuse behavior.	Yes/no; with Github, the data I am looking for are not findable because of the search engine; the search engine I use is easy to use for the type of data I work with but for other types of data (that I do not work with) search engines could be a problem; On Zenodo, I can never find what I am looking for	Yes ("The search engine on the open data repository that I use is sufficient for my open data search needs."): **2 respondents [I4, I8]** No ("The search engine on the open data repository that I use is not sufficient for my open data search needs."): **5 respondents [I2, I5, I7, I9, I10]**	The infrastructure's ability to contain a search engine that is sufficient for data search needs is an important factor for open data practices: **6 respondents [I2, I4, I5, I8, I9, I10]**	**Very High**

4.6. Case Study Description

This section describes the case that is studied in this master thesis project. The section highlights the important characteristics of Epidemiology concerning (open) research data sharing practices and the regulations that affect data sharing practices in the field. For references, the interviewee number is noted as the information source between square brackets (for example, [I1] refers to the first interviewee). To hide the gender of the participants, the third person singular pronoun "they" is used in the text when referring to the participants.

Epidemiology can be defined as "the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems" (Last, 2001, p.61). Thus, the field of Epidemiology is concerned about health-related phenomena in populations, and it is a method to find the causes of such

occurrences (CDC, 2012). Descriptive epidemiology is interested in distributions, it aims to answer the "what", "who", "where", and "when" questions in the five Ws journalism framework (CDC, 2013). Analytic epidemiology on the other hand is interested in the determinants, and it is interested in understanding the "why" and "how" of occurrences in populations (CDC, 2013).

For our case study, an important characteristic to mention about the field of Epidemiology is that an Epidemiologist may likely work in a clinical context. Clinical Epidemiology can be defined as a composition "between quantitative concepts used by epidemiologists to study disease in populations and decision-making in the individual case which is the daily fare of clinical medicine." (Last, 1988, p. 159). Clinical Epidemiologists are likely to have a medical degree, and as part of their clinic work, they provide patient care or treat patients. This is relevant for our case because doing clinical work has consequences in terms of time use and time resources that can be allocated for data sharing practices. [I3] brings attention to the difference between a researcher working in a clinical context versus one that does not: "*Where you [refers to the interviewer] do your research […], TU [Delft], has two activities, it's teaching and research. […] University Medical Centre has four tasks which is teaching undergrads, […] professional training, we have research and we have clinical practice. If you are a clinical epidemiologist, it's likely that a big part of that group is also a medical doctor. Being a clinical Epidemiologist, […] about 50% of us are also clinical doctors and the majority of those also have clinical duties. […]. The medical doctor among us has to juggle three tasks at least: clinics, research, and teaching. […]. [time] feasibility […] is more pronounced among clinicians than among other researchers because they have another task [refers to clinical work], and that task can never be postponed."* [I3]

4.6.1. The nature and source of Epidemiological research data

Epidemiologists may use primary data (i.e. the original data collected for a specific purpose by the researcher) or secondary data (i.e. data collected for another purpose by other individuals or organizations) in research activities (Mullner, n.d.). Apart from the secondary data that come from earlier studies (i.e., primary data of other researchers), other sources of secondary data in the field are patient medical records (hospital data), disease registries, insurance claim forms, mortality records, laboratory data, birth-death certificates, population surveys, environmental exposure data, public health department case reports and electronic health records (EHRs) (Hedberg & Maher, 2018; Mullner, n.d.).

As Epidemiology is interested in populations, researchers analyze and interpret large datasets in this field (CDC, 2013). This echoes in the common types of research study designs in Epidemiology. Four major types of Epidemiological research study designs are cross-sectional studies, case-control studies, cohort studies, and controlled clinical trials (Mullner, n.d.). For instance, "cohort studies" is a type of longitudinal study where certain research participants are followed over a long period, which is often years and sometimes decades (Barrett & Noble, 2019). Studying cohorts is found to be a common practice in this research field, and it has certain implications for data sharing practices. Because researchers have to obtain large-scale

datasets that contain information about individuals along a large time axis, data collection is extremely hard in these instances. Not only does it take a great deal of time to follow people for several decades and get measurements every couple of years, but it is also extremely costly to collect data in this manner (LaMorte, 2021). The fundings required for cohort studies are very large, and the duration of funding needs to be at least several years (*Clinical Trials and Cohort Studies Grants*, n.d.). Researchers who collect data in cohort studies do not always develop all the research questions at the beginning of this data collection. The research agenda could be flexible, and researchers develop new research questions and write new papers along the way. [I3] states, *"My studies are usually large and it costs a few million [Euros] or so to get the data. […] and sometimes 10 years follow up study, or 15 years. So I cannot specify all the research questions for myself before. Because you develop them as you are working. […] So it's [the research agenda] is flexible and develops over time"* [I3]. Regardless of whether the researcher is collecting data under a cohort study or not, Epidemiology researchers tend to use their primary datasets for several publications that answer several research questions. [I6] states, *"You may use your data to answer multiple research questions. When you publish an article, you probably […] only answer one research question"* [I6].

4.6.2. Repositories in use

Naturally, apart from researchers collecting data by themselves, many Epidemiology researchers use research data from other existing sources instead of collecting such datasets by themselves. In line with this, data sources such as "Biobanks" emerge as a type of resource that Epidemiology researchers can use for their research activities. A biobank can be defined as a structured collection of biological samples and associated data, which are stored for the purposes of present and future research (Parodi, 2015). Biobanks that contain large-scale datasets belonging to cohort studies are sources that Epidemiology researchers frequently use. One remarkable data source in the Netherlands is the Lifelines Biobank, which is a large and multi-generational cohort study that encompasses around 170,000 participants from the northern population of the Netherlands (*Over Lifelines*, n.d.). Since Lifelines Biobank is a not-for-profit organization, researchers are only charged for the price of the data release and the additional collection of data and biosamples (*Over Lifelines*, n.d.). Another Biobank that is commonly used by Epidemiology researchers is the UK Biobank, which is a biomedical database and research resource that covers a long-term biobank study in the United Kingdom.

Apart from the Biobanks, Epidemiology researchers may reuse or share research data on cross-disciplinary (i.e. not specific to any field) repositories such as Dataverse.nl (i.e. a publicly accessible data repository platform that enables sharing research data openly with anyone), Zenodo (i.e. an open repository maintained by CERN). Moreover, databanks such as Eurostat (i.e. statistical office of the European Union, that publishes Europe-wide statistics), StatLine (open data repository of Statistics Netherlands (CBS)). Some Epidemiology researchers also report having created their own websites/portals dedicated to data sharing. Some Epidemiology researchers also use Github to share the codes that they created during their studies (such as codes that are used to build mathematical models in infectious diseases, or codes that automatically extract data from other third-party open data sources). Environmental

Epidemiology researchers may use research data from open-source projects like OpenStreetMaps.

Regarding repositories that are more specific to their field, Epidemiology researchers use databases from World Health Organization such as VigiBase (WHO's drug safety data repository), and national databases such as Lareb (Netherlands drug safety data repository). Another repository that is reported in the case is MetaboLights, which is an open-access general-purpose repository for metabolomics studies and associated meta-data. Although not explicitly mentioned by the interviewees in our case, one noteworthy Epidemiology-specific open data repository is ClinEpiDB, which is an open access data repository that provides data from Epidemiological studies.

It shall be noted that some of the data repositories we report above do not fall into the definition of "open" data infrastructures, since researchers need to make an application to request the data, and (re)use of data is only possible if they are approved to do so. Lifelines Biobank, UK biobank, VigiBase, and Lareb are all data infrastructures that by default have explicit protocols to access data, which means researchers have to specify their research questions, and describe what the kind of contribution their research will bring to the fields, etc. if they want to use these infrastructures.

4.6.3. Data sharing in Epidemiology is (mostly) bounded by the privacy regulations

One of the reasons why Epidemiology or other similar fields that engage with human subjects have low levels of data sharing may be the privacy laws. The privacy law that concerns data sharing in the Netherlands is General Data Protection Regulation (GDPR), which is a regulation (that became enforceable in 2018) on privacy and data protection in the European Union as well as the European Economic Area (Hoofnagle et al., 2019). The GDPR applies to processing personal data in many contexts, one of which is research activities (Vlahou et al., 2021). While not stating any obligations for anonymous data (i.e. data from which no connection to an identifiable person can be drawn), the GDPR is concerned with the processing of personal data (i.e. Both directly identifiable and pseudonymized data) (Vlahou et al., 2021). The GDPR states that processing personal data is prohibited, and that an exemption can be made to this if consent is obtained (Vlahou et al., 2021). Therefore, getting consent is a legitimate basis for data processing as stated by the GDPR. According to the GDPR, the consent should specify the specific purposes (e.g. the objective of the research) (Vlahou et al., 2021). While the GPPR also prescribes an alternative way (an exception) of obtaining consent for purposes that are not specified in advance (i.e. obtaining "broad consent" for cases of scientific research while always respecting ethical standards), it is still unclear to research communities to what extent these exceptions could be legitimate (Vlahou et al., 2021).

Apart from obtaining informed consent (which is often the legitimate basis of data sharing in Epidemiology), another legal basis that can theoretically be followed in Epidemiology or other public health fields is the legitimate basis of "public interest", which justifies personal data processing for purposes to ensure the public's interest (Sullivan & Burger, 2017). The fields

working for public health indeed need to share data with each other to achieve purposeful research outcomes for the public, and this need for data sharing could be perceived as even more nuanced in Epidemiology since the field highly relies on the availability of large volumes of data.

Research data reuse is referred to as "secondary use" in privacy regulations, and this concerns reusing already collected datasets such as those collected in earlier studies (*Reuse of Already Collected Datasets (Secondary Use)*, n.d.). The GDPR does indeed allow for such further processing of data for academic research purposes because of the value given to academic research (*Reuse of Already Collected Datasets (Secondary Use)*, n.d.). However, The GDPR states that this is only possible strictly when this reuse is for research purposes and also when appropriate safeguards, technical and organizational measures, pseudonymization, etc. are taken by the organizations (*Reuse of Already Collected Datasets (Secondary Use)*, n.d.).

4.6.4. Data sharing practices in Epidemiology

Open research data sharing practices are currently not the priority in Epidemiology [I11]. The privacy regulations have recognizable influences on the lack of open data practices in Epidemiology. [I4] states, *"It is difficult to share [data] because of all the privacy issues [and] that you need to guarantee the privacy of your participants. […] People don't know how to do that [guaranteeing privacy]"* [I4]. In general, according to the GDPR, if researchers want to deposit the personal data (both directly identifiable and pseudonymized data) in a repository/archive (so it can be used by others), researchers must obtain the explicit, written, informed consent of the study participants (*Restricting Access to Data*, n.d.). However, this consent should specify to whom, for what use, and under what conditions the data will be shared (*Restricting Access to Data*, n.d.). This means that for the purpose of "open" data sharing, informed consent should indicate that the research data will be made public. When Epidemiology is concerned, getting informed consent from human subjects seems to be harder because data are often collected in the clinical context. [I7] states, *"In the case of clinical care [getting informed consent] is quite difficult because it's not something that we ask every patient ever."* [I7]. Since "open" data sharing is not included in the informed consent process, research data can be made public in these instances only if they are fully anonymized (where it leaves the scope of GDPR).

While data anonymization is, in theory, the solution to being able to make research data open without needing consent, for the field of Epidemiology or the field of healthcare, this may not be as straightforward. For example, [I3] states, "*In multicenter studies, sometimes it [the research topic] is kind of a rare disease. So in a sense, there are 10 patients in my data set, and if you have open data, then other people […] can say 'that should be my neighbor' who is in the database."* [I3]. Poulis et al. (2017) state that, concerning healthcare data, patient reidentification on a dataset (that is assumed to be anonymized) is in practice possible based on the patient demographics and diagnosis code, and that existing anonymization approaches are unable to prevent such reidentification attacks because data anonymization cannot ensure (1) that the data are still useful in the research tasks, (2) minimize the information loss, and (3)

fully guard against reidentification **at the same time.** Regarding this, [I11] states, *"I think data anonymization is a no go for epidemiological research or for a lot of health research in the sense that once you anonymize data and anonymize it in a way in which it is anonymized, per the GDPR, you lose so much value in the data. There is just no longer a point in it"* [I11].

When researchers want to make their research data open, they go into a process of privacy assessment in their organization, where research data management or privacy officers look into, among other criteria, the privacy risks of sharing the dataset in question [I11]. At this point, the request is likely rejected if the researcher had not obtained consent for data sharing when the data were collected. [I11] states that what often happens in practice is that researchers are recommended to share their research data with restricted access, since GDPR leaves room for the consent that was obtained for the primary study to be sufficient for reuse if this is for academic purposes and if strict organizational measures are taken [I11]. [I11] states, "*In order for a researcher to share epidemiological data, and let's say at a granular level, […] first of all there has to be a very strong legal basis. Often this will be consent […] Consent is for a specific purpose, and […] the GDPR does allow further processing for research purposes, which means that this data can be reused. But there must be organizational measures to make sure that if this data is to be reused, it is still being kept safe since it is still personal data"* [I1]. Therefore, researchers are advised to follow this data sharing practice over open data sharing to ensure compliance with the GDPR. By sharing data with restricted access, there is some level of control on who the data are shared with and what these people will do with the data. After all, if the data are made fully open, then there is no possibility to control who would use the data and for what reasons the data would be used. [I11] states, "*What is often done or recommended is to share the data on the restricted access so that at the very least they can have a data transfer agreement or some control over whom they are sharing the data with, because once you make something publicly available, you also cannot guarantee who's going to be using this data"* [I11]. If data are shared publicly, it is not possible to know whether the data will end in a very authoritarian country or somewhere where the individuals will be reidentified for profit and commercial interests [I11]. While it is not possible to prevent such outcomes if the dataset is publicly available, the risk can be somewhat minimized if data are shared with restricted access.

A common data sharing practice in the field of Epidemiology is data sharing by request (or, in other words, one-on-one data sharing), which refers to sharing of data when the requester party expresses interest in the dataset and asks for approval for use. Researchers mention in their publications that the corresponding research data would be available within reasonable request: "*We mention in our work that the data are available upon the reasonable request, so it is not open out there on the internet. But if researchers want to use it, they could contact us"* [I1]. [I3] states, "*I do get emails from people I know somewhere in Europe or in the Netherlands […], who ask me, 'I see your publications and I see you have nice data […], could we collaborate?' Then I share my data"* [I3]. It is also possible the data requesters are invited by the researcher who holds the primary data to come and understand the dataset together, so that the requester is given all the necessary information to properly interpret the dataset. [I3] states, "*Usually I ask them [the data requester] to come over for two or three days and to do here the*

analysis because the database is quite complicated and then it is easier for them to sit here and talk to my PhD students and they help them to find the right variables and so on […] It is about interpretation because I don't want my study to be wrongly analyzed by other people" [I3].

Another way of data sharing is data sharing by collaboration, which is when parties share research data and collaborate to produce a joint publication, which means that the researcher who held the data in the first place gets credit in the publication at the end. Data sharing by collaboration can take many forms depending on the level of collaboration (i.e. solely research data sharing or data sharing along with conducting a study). There could be less formal collaborations, for example, in multicenter cohort studies where researchers from participating institutions deliver research data to a research team that is conducting the study. The research team who receives the data gives the participating organizations a chance to collaborate if they are interested: [I3] states *"[We tell them] 'you'll deliver data, we don't pay you for it, and if you have your own research questions or ideas, please contact us and we will help you to validate [them] or do [the research] together'. So, we do collaborate in this way"* [I3].

4.6.5. Lack of an open data sharing culture in Epidemiology

Compared to fields that do not engage with human subjects, the level of "data sharing" practices in the field of Epidemiology is perceived to be lower. [I11] states, *"If you compare it [data sharing in Epidemiology] to the geosciences or physics that it is quite low simply because those are […] hugely open fields"* [I11]. However, due to its nature of examining "populations", data sharing in the field of Epidemiology could be higher than in other life sciences fields. [I11] states, "*If you compare it to, let's say, neuroscience or compare it to a lot of life sciences […], then I believe that there is much larger propensity for these sort of data [Epidemiological data] to be made available simply because Epidemiology is about […] acquiring as much data as possible in order to get good comprehension of the population"* [I11]. In that regard, the field of Epidemiology is not unwelcoming to data sharing by forms of collaboration where all parties get acknowledgment in the outcomes of research. [I6] states, "*If it is a collaboration, it is possible […] for researchers to share their data. […] sometimes when you share a common interest it is possible for you to collaborate and in that sense data sharing is possible in a collaborative way*." [I6].

Regarding "open" data sharing, researchers consider the field to lack a culture of open data sharing. [I3] states, *"There is certainly a culture, but that not a fully open data sharing [culture], but it is data sharing. I really appreciate to collaborate with these people, and people contacting me, and I approach other people in congresses, and so on. So we do a lot of data sharing."* [I3]. [I1] states, *"culture-wise it [having an open data sharing culture] is more difficult […] it is a bit individualistic and people are very focused on their own track record and publications"* [I1]. [I4] states, *"[there are] large cohorts of patients and, people can be quite protective of their data*" [I4]. [I6] states, *"[…] I do not believe that you will get any data from other people without collaborating with them."* [I6]. [I7] also states that even within the organization people are not open to data sharing practices if they would not get an acknowledgment (e.g. via publication).

4.7. Case Study Analysis

This subsection first presents how infrastructural and institutional instruments for (open) data sharing and reuse function, how they are perceived, and what roles they take in Epidemiology as identified by our case study. Then it explains the leading barriers in front of open research data adoption as indicated by the participants. Furthermore, it examines the relationship between the instruments and presents a refined conceptual model based on the case study analysis. Finally, it summarizes the key characteristics of the Epidemiology field that concerns open research data adoption. As in the previous chapter, for references, the interviewee number is noted as the information source between square brackets (for example, [I1] refers to the first interviewee), and to hide the gender of the participants, the third person singular pronoun "they" is used in the text when referring to the participants.

4.7.1. The role of Infrastructural instruments in Epidemiology

Table 9 Infrastructural instruments to support open research data sharing and reuse (identified through the case study) (see chapter 4.7.5. for a visual illustration)

Infrastructural instrument category	**Identified infrastructural instruments**	**Availability of the instrument** *(See chapter 4.5.4.1. for operationalization)*		**Importance of the instrument** *(See chapter 4.5.4.2. for operationalization)*		**Level of influence on data sharing and reuse in our case** *(See chapter 4.5.4.3. for operationalization)*
		# of respondents who say they use the instrument	**# of respondents who say they do not use the instrument**	**# of respondents who find the instrument important for open data practices**	**# of respondents who find the instrument unimportant for open data practices**	
Instruments enhancing the usability of infrastructures	Easy to use, convenient interfaces	5	2	4	0	High
	Capability to handle a large volume of data	5	2	3	1	Medium
	Compatible and/or integrated infrastructures	3	5	5	1	High
	The data repository allows for data analysis (integration)	2	3	1	2	Low
	Availability of data management tools	4	1	1	0	Low
Instruments supporting the facilitation of	Availability of a search engine that is sufficient for open data search needs	2	5	6	0	Very high

FAIR data principles	Availability of higher-level search engines/registry of repositories that enable researchers to search data across different repositories	0	5	4	0	High
	Availability of data usage statistics on the platform	1	3	0	3	Low
	The infrastructure offers metadata	6	1	2	0	Medium
	The infrastructure offers metadata on data collection methods	-	-	4	0	High
	Availability of tools that are used for metadata creation and management.	0	2	1	0	Low
	Offering assistance for data citation	2	0	-	-	Unclear (influence on open data practices is not discussed)
	The infrastructure is compatible with domain-specific privacy requirements	0	2	3	0	High
Instruments concerning security and trust aspects	Availability of data anonymization tools	1	4	0	3	Low
	Offering ways/methods to assess how trustworthy an open data repository or an open data set is.	0	3	-	-	Unclear (influence on open data practices is not discussed)
New instruments detected through the case study	Fast download process	-	-	2	0	Unclear (instrument brought up by only two respondents)
	Standardize way of working among different repositories	-	-	1	0	Unclear (instrument brought up by only one respondent)

	Enhance the usage of unique identifiers, such as the ORCID identifiers	-	-	1	0	Unclear (instrument brought up by only one respondent)

4.7.1.1. Instruments enhancing the usability of infrastructures

Easy to use open data infrastructures

Many participants who reused datasets from repositories state that the (graphic) interfaces of the data infrastructures such as the data repositories that they used are user-friendly and convenient to use. For data reuse, it seems that a user does not have much engagement with the platform, which means that oftentimes the process of extracting data does not cause problems for the researchers in terms of interfaces [I7, I2]. [I2] states that data reuse is just a matter of getting a CSV file from the repository or through an API, so there is no issue for ease of use. Regarding data sharing, on the other hand, the situation may be different because several interviewees note the instrument as an important factor for open data practices. [I9] and [I10] state that they had issues in the past with dealing with the interfaces of the open data repositories, both reporting the issue of not being able to see/find the dataset they uploaded to the data repository. [I9] even state that the issue of the complicatedness of the repository almost inhibited them from uploading their data on the open data repository: "*I once used open data source MetaboLights. A journal wanted me to upload all my data […] It was complicated. […] It kind of inhibited me from doing it [sharing the data openly] almost. […] I didn't understand [how to upload the data] […] Then it said I had successfully uploaded. I couldn't find it. I couldn't find my own data."* [I9]. [I7] mentions that especially with Github, something that is often not realized is that researchers may need trainings to get comfortable with using the infrastructure: "*I guess [Github] is relatively easy, but it has a very steep learning curve to get there. […] You really need some training in how it works and also how it works for this specific group that you're working with. This is not something that is often realized by researchers. They just tell you 'there is GitHub and go there' and you're like, 'well, I don't really know what I'm supposed to do there'. […] I think that would really be something more as an awareness point that people should be more aware of how to train their new team members in using this kind of infrastructures."* [I7].

Capability to handle a large volume of data

Five of the interviewees state that the repository they use can accommodate as much volume as they need for their research purposes. Since the infrastructure's capability to handle a large volume of data is perceived as an important factor by some of the interviewees, the instrument is found to have a medium influence on our case. [I1] states that their research would not have been possible if the infrastructure did not accommodate large-scale data. However, it is unclear whether this functionality is always an important factor for open data practices for all research types or subfields in Epidemiology. The type of subfield and data the researchers work with influences the volume that researchers require from the data infrastructures, and this may influence how they evaluate the infrastructure's capability to store large amounts of data. For example, [I2] states that the volume capability is not much of an issue for the type of data they

use for their research. [I10] states that some subfields of Epidemiology, such as the genetic-related subfields, require large volumes of data, which means the infrastructures would definitely need to accommodate a large volume of data in certain studies to facilitate open data practices: *"For my need, so far it's [the volume capability of the infrastructure] been large enough, but that's also because it didn't include yet genetics data. […] If I would do that, I'm not sure if it [the volume capability of the infrastructure] would be large enough. […] Some Epidemiological studies have very extensive laboratory results including genetics data […] I can see that there could be studies where you would definitely need this [functionality]. […] For instance, studies where you monitor patients in the hospital and you also store all the data from the monitoring systems like your heart beats and temperature [continuously]. […] That's again a lot of data points and there you also need very large storage. […] It depends on the type of research, I would say."* [I10].

Compatibility and integration between infrastructures

The compatibility and integration between different data structures are perceived as an important factor for open data practices by many researchers [I1, I5, I7, I8, I9, I10]. From the statements given by the interviewees, there is an indication that currently, the level of compatibility and especially the integration between infrastructures that are used for open data practices are not at satisfactory levels. [I8] and [I4] state that (full) integration is currently not a reality, but, for [I8], compatibility is an essential element for open data practices. There is evidence that these integrations ease the process of data sharing or data reuse of researchers. [I2] gives an example of some infrastructural integrations that they perceive to be helpful, such as how Open Science Framework (OSF) is well integrated with Github, with storage applications like Dropbox, and with discovery applications like Google Scholar and ORCID. [I7] also mentions the integration between Zenodo and Github. Several interviewees express their dissatisfaction regarding the compatibility between different infrastructures, especially compatibility issues due to data types. [I10] gives an example of a compatibility issue, and states that when they take straightforward datasets from repositories, the files are mostly in compatible formats, but when they take datasets from different APIs of applications or other internet interfaces, then compatibility could be a greater issue: "*Sometimes we struggle because so we work with R most of the time, and some data sets are easy to upload into your environment, like a .CSV file […]. But nowadays more and more data that is collected through an internet interface like apps -we use apps to monitor persons, etcetera-, they come in XML files, and that is more complicated. […] So the more straightforward the data set is, […] that's very easy, but if it gets more layers, […] then then then getting it in into your R environment in the right format can be quite a challenge. […]"* [I10]. When asked if this issue poses a strong negative influence on open data practices in the field, [I10] adds, *"I do think so because […] a lot of this type of work -where people would start working with [reusing] data sets from somewhere-, is done by PhD students. Now, you almost need to have a background in computer science to be able to deal with all the different types of data sets and to get that in your statistical analysis software environment. […] So it requires a new set of skills […] which is not typically what you learn in your masters [in] Epidemiology"* [I10]. This indicates that there could be a relationship between certain compatibility issues and the need for training that researchers need to receive to be able to engage in open data practices.

Furthermore, [I7] mentions that because of privacy protocols, some infrastructures purposefully restrict integration with other infrastructures when retrieving or uploading data, and in such instances, data sharing gets really hard for the researcher. Although the following example is not about open data infrastructures, it still indicates the value of being able to use technical infrastructures (e.g. software, databases, the internet, etc.) without conflict when engaging in data practices. [I7] states, *"I linked my data set that I had to the Dutch National Statistics. […] But then again, you could only analyze stuff in the environment and the environment was completely shut off from anything. So even the Internet wouldn't work on your computer when you were in that environment. […] So then again, […] anything you wanna get out of that environment and share, people have to approve for you. So that makes it really hard to share."* [I7].

Facilitating data analysis as an integrated feature

The feature of facilitating data analysis -as an integrated feature- on the data platforms has a low influence on open data practices in our case since this instrument is not found to be an important factor by interviewees except for [I1]. [I1] mentions their satisfaction with being able to perform data analysis integrated into Lifelines Biobank's workspace. However, other interviewees do not label this feature as an important factor for open data practices. There is also evidence that currently, open data platforms do not have such integrated features. However, our analysis suggests that researchers do not need such functionality. In fact, [I5] and [I9] state that the open data repositories they engage with do not allow for an integrated data analysis activity, and that there is actually no such need. [I9] states *"In the end, researchers want just to get the data […] Do they really want to depend on some analysis tool that is preprescribed in a [data repository]? I don't think many researchers want that"* [I9].

Availability of data management tools

Several researchers report using management tools available to them as part of their regular research activities. The most notable mention is that of DMPonline, which most interviewees seem to have used for creating data management plans. [I7] mentions using the digital research environment AnDREa, which helps researchers in storing, analyzing, and sharing data. However, our analysis indicates a low influence of the availability of these tools on open data practices, as none of the interviewees (except for [I2]) talks about whether such tools facilitate their motivation toward open data practices. [I2] states that having access to data management tools for managing the plan and managing the actual data (including the data collection) is valuable for open data practices. [I4] gives an example of a data management tool that is used for internal data sharing (among the organization). Furthermore, [I8] states that oftentimes researchers need to find data management tools (apart from Dmponline) by themselves, so the organization does not support them by offering such tools.

4.7.1.2. Instruments supporting the facilitation of FAIR data principles

Availability of a search engine that is sufficient for open data search needs

Offering a powerful search engine tool is found to be a strong factor in open data practices by several interviewees. Arguably one of the most discussed-about instruments during the interviews was the (lack of) well-functioning search engines on the data platforms (repositories). The majority of the interviewees state their dissatisfaction with the current search engines on the repositories, citing major problems regarding the findability of research data. [I8] states that the functionality of the search engines is very important for open data practices. [I2] states that in their experience finding data has been a major problem, and thinks that not being able to find the data is definitely a barrier for open data practices. [I4] also states that it is likely for a researcher to experience issues in searching for a dataset on CBS since it is complex, and that this issue could demotivate researchers from data practices: "*It's hard to get it your way in there [CBS]. […] I can imagine […] it would be an issue to find what you want there, so I think this indeed can hinder [some researchers'] motivation" [I4].* If you know what you are looking for (i.e. if you have a DOI at hand) then reaching the data is easy, but if you have to "search" for data, then there are difficulties because the search engines are not working properly (i.e. there is no particular way to search for a specific topic under the field) [I2]. During the interview [I2] even demonstrates an exemplary search on the open data repository Zenodo, where they type an Epidemiology related keyword, and shows that many of the results are results that are completely unrelated to the search (most results are what [I2] calls "junk", which inhibits motivation for open data reuse). [I7] also cites the same problem with the Zenodo platform: *"I can never find the stuff I need on it."* [I7]. [I9] mentions the problem of insufficient search engines as the most troublesome aspect of open data infrastructures such as GitHub, citing the same reasons: If you are reading a paper and if you would like to access the data that are linked on that paper, then it is very easy to reach the data on such platforms. But if you want to find data or summary statistics on some specific topic, findability is a problem. "*The difficulty about GitHub is, I don't think it's [the data] very findable. […] [If] you're reading a paper, then […] you click the GitHub link and then you find it. But if you say 'I want to find summary statistics or I want to find data on Diabetes in […] neighborhoods in the Netherlands', how do I find that? […] That's maybe the most difficult thing is that […] we're very accustomed to PubMed or Google right to find articles or find information. But for finding data, a proper search engine for finding data that relates to your question, is there a search engine like that? Do you know of a search engine that does that? […] So that makes it really, actually unfindable."* [I9]. Indeed, researchers are very accustomed to performing search queries on Pubmed or Google to find papers, but when it comes to searching for data, they seem to be struggling. [I9] and [I10] wish to see a well-functioning search engine that is similar to the one on Pubmed, where you can search for data with "extensive search options". [I10] states, "*You have these large online searchable databases like PubMed or medical archives where you can have a very extensive search strategy. […] And that really helps to find what is relevant for your work. And I don't think that [in] the field of databases it's so much developed yet."* [I10]. This is also fully in line with the literature findings from Behnke & Staiger (2019) who state that it is essential to provide various query interfaces to accommodate different data search behaviors for the search engines.

Availability of higher-level search engines/registry of repositories that enable researchers to search data across different repositories

Similar to the issue described above is the issue surrounding (the lack of). aggregating/overarching registry of repositories (or a search engine) that Epidemiology researchers can make use of. Interviewees state that the individual repositories that they use are not linked to any aggregator infrastructures where it would be possible to search for data across resources [I1, I2, I5, I7]. [I2] states, *"There's no way to search for everything [meaning across repositories]"* [I2]. Furthermore, they express the need for such an aggregating search tool and believe that this is an important factor for open data practices. [I9] states that researchers should not have to go to Google (databases) when they want to find a dataset belonging to a certain demographic in a specific region. Instead, there should be a search engine available for this (which [I10] calls, for example, "a PubMed for research data"), to where all the data repositories are linked, and this infrastructure should print the datasets linked to your keywords or your extensive search queries [I5, I9, I10]. [I9] states, *"If I'm a researcher and I want to do a search on [a certain medical practice in a certain geographical location] related to social demographic factors, I should not go to Google, but that's still where people start. But that's not where you should be starting. You should start in some search engine that says 'I have all data repositories linked to me, and I will find you different data sets linked to your keywords.' […] But I haven't found [such a search engine]."* [I9]. [I11] states, *"I think that there's not a good open research data search engine yet. You have Google databases, you have a few [engines] here and there, but, for example, one of the famous repository search engines where you look for repositories, it is abysmal. I do not recommend it. Half of the links are broken. I think that there is definitely a niche or a spot there to be filled by a proper research data search engine."* [I11]. The availability of an overarching search engine/registry is found to have a high influence on open data practices in our case.

Availability of data usage statistics on the platform

Regarding the availability of data usage statistics on data repositories, some interviewees state that the repositories they use do not show such statistics [I1, I5, I8]. [I7] states that repositories like UK Biobank do show such statistics or in some other way the repositories at least present some kind of information on whether other researchers published on the particular dataset and what kind of topics were published on such datasets. However, [I7] states that they have never specifically looked at such information themselves. This is also in line with other researchers, such as [I8], explicitly stating that they are doubtful whether anything would change in terms of open data practices if such statistics existed on the platforms. None of the interviews cited data usage statistics as having important potential for being a promoter for open data practices, which is why the influence of this instrument seems to be weak in our case.

The infrastructure offers metadata

Most interviewees state that the infrastructures they use preview metadata to some extent [I1, I2, I4, I8, I9, I10]. [I8] and [I7] state that being able to find metadata is a very important factor in open data practices. [I7] states, *"It makes me very happy [to have metadata]. […] the data set I used for my PhD research […] was just CSV files, we got none, no metadata at all. So*

that was just me for like half a year puzzling, trying to figure out what all the variables mean. […]. So I'm very happy when I get proper metadata." [I7]. Our analysis shows that generally, researchers do not have trouble accessing metadata for the datasets they reuse. However, metadata is an umbrella term. The current problem with metadata on these platforms could be about their content rather than their mere existence, which we discuss in the next subsection.

The infrastructure offers metadata (or data dictionary) on data collection methods

Although many interviewees state that it is currently possible to find metadata on the open data infrastructures, the current problem with metadata on these platforms could be about their content rather than their mere existence. This relates to the instrument of showing metadata (or a data dictionary) specifically on data collection. [I1] states that being able to see how each variable was measured and what was exactly asked (when the data were gathered) is very important. When discussing their data sharing practices, [I2] states that rather than making their dataset fully open, they would prefer to invite the data requesters to visit them for two or three days, because the database may be complicated; by doing this, it is possible to inform data requesters on the variables. [I2] adds that they do this because of interpretation: they do not want their study to be wrongly analyzed by others because of an incomplete understanding of the data. [I5] also brings attention to how the lack of such metadata may be an important barrier to open data practices: *"[If] you're not too sure how data is collected… These kind of things [demotivate]. What that variable means? If it's measured with this instrument or with another instrument…"* [I5]. [I9] mentions the same issue about metadata on data collection: "*If somebody says 'this weight of a person', was it measured, was it self-reported? If it's not mentioned, then how am I supposed to know how it was measured? How did they assess the weight? That's the problem with data dictionaries, they don't [do that]. Some are very limited. So, then you have to go back to the researchers [who prepared the data]. […] It is difficult"* [I9]. The instrument is found to have a high influence on open data practices in our case. Our analysis indicates that currently, the data dictionaries or the associated metadata in Epidemiology do not sufficiently provide details about the data, considering that researchers say it is very important to know all the properties of the data down to every fine detail. If these details are not explicitly mentioned along with the dataset, then there is no way a person who would like to reuse the data can understand them.

Availability of tools that are used for metadata creation and management

Although being able to find a description of the data with sufficient quality and depth seems to be an important element for open data practices, the interviewees do not find the availability of tools for metadata creation and management to be an important factor for open data practices. It seems that researchers do not use dedicated tools for metadata creation but create metadata with other available tools [I7, I8]. [I8] states that they may be interested in having access to such tools, but does not mention how this instrument can affect open data practices. Apart from this, none of the interviews cited being able to access or use metadata creation or management tools as having important potential for being a promoter for open data practices, which is why the influence of this instrument seems to be weak in our case.

Offering assistance for data citation

The relationship between the infrastructure's ability to offer assistance for citation and open data practices seems to remain unclear in our case. [I1] states that they do get help from the Lifelines Biobank infrastructure on how to properly cite the research data. [I5] also states that the infrastructure that they use describes how to properly cite the research data. However, interviewees do not talk about whether assistance in data citation has any relationship with open data practices, which is why our analysis is not conclusive in this regard.

The infrastructure is compatible with domain-specific privacy requirements

The domain of healthcare strictly requires the researchers to abide by privacy regulations when they are engaging with open research data practices. Many researchers noted privacy regulations as one of the strongest barriers to open data practices (see chapter 4.7.3). Relating to this, the open data infrastructure's compatibility with (domain-specific) privacy requirements (as an instrument) is also found to have some influence on open data practices. [I7] states that the existence of privacy rules leads to the necessity of these technical systems to gain new features that enable the accommodation of sensitive data (without violating privacy rules). [I9] brings attention to the fact that it is currently really hard to link datasets to overarching registries because the data repositories do not give any easy space to deal with privacy regulations, which suggests that researchers are somewhat expecting support from infrastructures on how to overcome issues stemming from privacy regulations. [I7] states, *"It is quite hard actually for us in this field to share data because it's often patient level. So I do feel like especially for medical data sets -patient level data sets- that if we would want that to be more open, you would need some kind of [an] infrastructure. I don't think any of the infrastructures that are out there right now cater to this kind of data and therefore, when you talk about sharing your data in a repository or online or anything, everybody just tells you can't really do it. […] People are just saying no because of the privacy rules. […] And I think sometimes that's a bit of a shame, because there might be actually ways to work around it, but there isn't really anything [any infrastructure] facilitating that at the moment."* [I7] Interestingly, some interviewees provide examples of infrastructures and concepts that can address these concerns. One such example is the OpenSafely initiative. OpenSafely infrastructure (although not a fully open data platform by definition) allows researchers to access sensitive data without breaching privacy (*About OpenSAFELY*, n.d.). Researchers can analyze data without actually accessing the data themselves. Researchers use dummy data for developing their analytic code on their local computer and using the code, they can perform the analysis on the data, without ever accessing the data (that always stays in the secure environment) (*About OpenSAFELY*, n.d.). [I11] also talks about a similar concept: *"If you [a researcher that wants to work with a certain dataset] have a particular analysis that you want to do on the variables [that you are interested in], then you can send the analysis to the people that currently control the data. They can do the analysis and give you back an aggregated result, which is then anonymized. […] You can automate this to some degree."* [I11]. Infrastructures like OpenSafely safeguard against important privacy issues, because patient-level data are never seen. This functionality can be very important for data practices in Epidemiology. In line with this discussion, the infrastructure's compatibility with privacy requirements gains importance. [I11] gives another example of an infrastructure that tackles

the problem of privacy, which is the synthetic data solution: "*The way this works is that you basically create a statistically similar population. […] There's a statistical mimic of the original data set. So that means that statistical tests on this new data will give you very similar results. [..] Of course, this doesn't really work for testing hypotheses, it's more for feeding machine learning algorithms, which is something that is starting to be done quite often. And I think for that particular purpose synthetic data can be quite useful."* [I11]. Our analysis shows that there is a need for reviewing methods like this, for example, to see whether synthetic data generated via an AI engine that simulates an existing dataset with identical correlations and patterns can in practice benefit researchers in Epidemiology (Lin, 2019).

4.7.1.3. Instruments concerning security and trust aspects

Offering ways/methods to assess how trustworthy an open data repository or an open dataset is

As the literature suggests that trust that researchers attain to open data repositories and open datasets could be a factor in open data practices, the instrument of offering ways to assess how trustworthy an open data repository is discussed with the interviewees. The researchers do not have access to any tools or methods to assess how trustworthy an open data infrastructure is [1, 4, 7]. It is also noteworthy that several interviewees state that they trust the data repositories they use [1, 5, 7, 8, 9] or that they have not considered before whether the repository they use is indeed trustworthy or not [I4], which could indicate that researchers do not experience issues in maintaining or building trust towards the repositories in open data practices. No interviewee reported any connection between having ways to assess the trustworthiness of a data and their behavior towards open data practices, which is why the influence of this instrument is marked as low.

Availability of data anonymization tools

In theory, data anonymization could be a valuable tool for open data practices in the field of Epidemiology since it tackles issues stemming from privacy regulations. However, when data anonymization tools -as an instrument- are discussed with the interviewees in our case, the instrument is found to have low influence on open data practices. It seems that most researchers do not use exclusive software for data anonymization [I4, I7, I8, I9] and more importantly, express that there is no need for it. [I8] states that anonymization may be important for open data practices, but they do not think specific tools are needed for anonymization, and that researchers can do this without using exclusive software. Similarly, [I9] states that researchers do not struggle with data anonymization and they are usually able to perform this act by themselves. In chapter 4.6.4, we discussed that data anonymization might not be as beneficial for the field of Epidemiology (or the fields of public health) since full anonymization makes data lose value significantly (Poulis et al., 2017). Therefore, our earlier findings on data anonymization methods not being entirely useful could also be a reason why the data anonymization tools are not cited as important tools by the interviewees in our case.

4.7.1.4. New instruments that emerged through the case study

Fast download process

Relating to this discussion of infrastructure's capability to handle a large volume of data, a new instrument that emerged during the interviews is the data download speed of the open data infrastructures. Some researchers expressed dissatisfaction with the speed with which they downloaded data from repositories. [I4] states that because they work with a large volume of data, the infrastructure can only provide the data by disaggregating them into smaller pieces, which costs a lot of time and effort. [I7] also has had a similar issue, where a download for a certain dataset took two weeks. [I7] brings up the possibility to switch to cloud-based technologies for these infrastructures, but also states the issue of privacy could be a barrier to such a switch. [I4] states that this problem (of needing to put a lot of time to download huge datasets) is not the strongest barrier to open data practices, but it is definitely an issue that interferes with the user experience.

Standardize way of working among different repositories

Another instrument that was brought by the interviewees is the standardization of working among different repositories [I8]. [I8] states that individual repositories are very rigid and strict: a user not only has to adhere to the lingo and the language of the repository, but they also always have to adjust and reformulate everything for each different repository. [I8] believes that there should be a standard way of engaging with a repository as a data depositor or data requester. As this concept was only mentioned once, it is not yet clear how much this functionality influences open data practices. The number of repositories Epidemiologists work with is indeed high, which is why examining this instrument further could be valuable.

Enhance the usage of unique identifiers, such as the ORCID identifiers

The final instrument that emerged through the interviews is enhancing the usage of unique identifiers (such as the ORCID identifiers). An ORCID (Open Researcher and Contributor ID) identifier is a 16-digit number that helps researchers to distinguish themselves from other researchers, by enabling them to connect their identity (name) to their research publications and professional activities. The benefit of having an ORCID identifier for a researcher is making sure that the researcher's work always has visibility, which relates to the possibility of getting credit for the work. [I2] states that thanks to having an ORCID identifier and to the identifier's easy integration to repositories like GitHub, their data contribution to the repository is well connected to their name, and that they will sufficiently get credit for their work when other people use their data. [I2] states that their organization requires them to link their work to their name using the ORCID identifier.

As the concept of identifiers and their relation to open data practices (through ensuring proper credit) was only mentioned once, it is not clear how much this functionality influences open data practices in our case study.

4.7.2. The role of institutional instruments in Epidemiology

Table 10 Institutional instruments to support open research data sharing and reuse (identified through the case study) (see chapter 4.7.5. for a visual illustration)

Institutional instrument category	**Identified institutional instruments**	**Availability of the instrument** *(See chapter 4.5.4.1. for operationalization)*		**Importance of the instrument** *(See chapter 4.5.4.2. for operationalization)*		**Level of influence on data sharing and reuse in our case** *(See chapter 4.5.4.3. for operationalization)*
		# of respondents who say they use the instrument	**# of respondents who say they do not use the instrument**	**# of respondents who find the instrument important for open data practices**	**# of respondents who find the instrument unimportant for open data practices**	
Instruments that manage and govern data sharing and use process	Offering institutional data sharing policy and guidelines for openly sharing and reusing research data	8	3	1	4	Low
	Offering institutional data management policies	7	1	1	1	Low
	Asking for data management plans	9	0	2	3	Low
Instruments that actively support researchers in sharing and using research data	Support from data stewards	3	4	5	0	Very high
	Working with research data managers	5	5	5	0	Very high
	Enhance the library's role	3	4	3	1	Medium
	Training and educational support	5	1	2	0	Medium
	Providing support for legal aspects (privacy) of open data practices	5	1	4	0	High
Instruments that relate to financial resources	Providing separate funds for research data management	0	5	2	0	Medium
Instruments that build a culture of data sharing and create incentives	Recognizing and rewarding open research data sharing contributions	0	6	4	0	High
	Considering data sharing contributions during hiring, tenure, or promotion decisions	0	6	2	0	Medium

	Requests for open research data sharing from organizations, funders, journals	10	0	1	1	Low
	Demonstration of benefits of data sharing and the need for data sharing	0	2	2	0	Medium
	Clarifying the concept of data ownership	5	1	2	0	Medium
New instruments detected through the case study	Building an open science community within the Epidemiology field	-	-	2	0	Unclear (instrument brought up by only two respondents)
	Increase communication among the scientific community so that two people with similar research interests are aware of each other	-	-	1	0	Unclear (instrument brought up by only one respondent)

4.7.2.1. Instruments that manage and govern data sharing and use process

Institutional data sharing policies

Most interviewees state that there are institutional data sharing policies in their organization. The data sharing policies in these organizations focus on guiding researchers in a variety of aspects ranging from how they should share data, how they should make a data management plan, and how they should make their data as open as possible [I11]. For example, the website of UMC Utrecht mentions that the research data should be under the FAIR principles (except for when it legally cannot be). It also thoroughly guides researchers about which domain-specific repositories can be used, and under what conditions research data can be reused or openly shared. However, several interviewees state explicitly that, although there may be (open) data sharing policies or guidelines in their organization, they do not know much in-depth about the content of these policies [I7, I9]. Some of the interviewees mention that they are not aware of such policies [I1, I4, I5]. However, when the websites of some of the organizations, where the interviewees work, are analyzed, data sharing policies and guidelines can actually be easily accessed (at least internally), which signals a problem that researchers are not engaging with policy documents as intended by the policymakers. Regardless, all participants are well aware of the FAIR principles in the open science paradigm in our case.

Our analysis shows policies do not have a high influence on open data practices. There seems to be a misalignment between what the policy of the organization states on paper versus what the researchers do in practice. Researchers indeed may feel as if there is no need for going in-depth into the content of policies. For instance, [I7] states that they never felt the need to look at these documents. [I11] states, "*I don't think they're very useful. Most people don't read them. […] there are people that read them because they wanted to share their data to begin with. So, they probably would have [read] anyhow. […] I don't think they're [policies] having the impact that we hope they would. They're mostly just words on paper. […] I strongly believe policies*

[…] have to be more than just words on paper. There has to be a way in which this [policy] is relevant to the researchers, and that making data available […] is beneficial to them, not just because they're following a policy." [I11]. [I2] states that the policy in their organization states that the data should be FAIR, but that the policy on the institutional level does not reach the "work floor": *"So policy on institutional level might not reach the work floor. […] In principle, our data is FAIR. That's the policy. The reality is that the one that is responsible [for this principle] is 3 levels down [than the policymaker]. […] It is the head of department that is responsible, and not only responsible, [but] also free to choose. […] And the moment that people don't adhere to this [principle], there will be a moment where somebody asks, 'hey, why are you not doing this?'. But there are no […] repercussions both formally and informally. […] the policy […] is very difficult to enforce."* [I2].

Researchers may also hold opinions that, to some extent, contradict the "make data as open as possible" principle in the policy documents. [I3] states that they are aware of the statements of the organization's policy (regarding the FAIR principles), but then explains why they do not prefer making the research data fully open and why they instead favor data sharing upon request. [I9] also explains that despite the policy, researchers have the discretion not to share data in their organization. Considering that the policy documents have a weak influence on open data practices, one could suggest making policies more enforceable so that the expected effect can be realized. However, the interviewees also express that policies should not enforce but rather take the role of an advisor on these topics [I3, I9].

Institutional data management policies

Regarding the institutional data management policies, most interviewees state that they are well aware of these policies and that they believe these policies hold value -to some extent- for research activities [I4, I5, I6, I7, I9, I10, I10, I11]. When the policies are reviewed by us, it seems that policies indeed give clear guidance on various research data management topics. For example, when UMC Utrecht's data management policies are reviewed, it can be seen that the document gives extensive information on how researchers should prepare for the research (e.g. how to create a data management plan, how to store files), collect data (e.g. how to reuse data from previous studies), prepare data (e.g. how to anonymize data), create metadata, analyze the data, archive the data and share the data. [I6] states that they are well aware of how to make an application to use an existing dataset for their research due to the data management policies. [I5] states that sometimes following these policies/guidelines is too burdensome. While it is clear that researchers have access to data management policies, and that these policies guide various data sharing activities, our analysis does not point to a strong relationship between institutional data management policies and open data practices. Interviewees mostly do not explain the relationship between data management policies/guidelines and open data practices. Therefore, this instrument is indicated as one that has a low level of influence on open data practices in our case.

Asking for data management plans

One of the related instruments under institutional data management is the data management plans, which are also promoted by policy documents in the organizations. The majority of the

interviewees state that they make data management plans for their research (although not for all studies). Organizations ask for data management plans as part of the research cycle while it is usually not a mandatory requirement [I2]. Some interviewees state that funders require data management plans [I7, I8, I9] or that making a data management plan is a necessary step when you are applying for grants [I4]. From the discussions, it is clear that researchers are well aware of the function of these plans, and the steps needed to conduct to make a data management plan.

Regarding whether the instrument of data management plans influences open research data practices, our analysis points out a clear pattern in the level of open research data experience that a researcher has and the attitude toward data management plans. Researchers with low experience in open data practices tend to find data management plans useful and do not express negative feelings towards them. [I1] (who relatively has a lower level of open research data experiences) believes that thinking about how to deal with data before the research starts helps them work in a structured way, and [I4] (who also relatively has a lower level of open research data experiences) believes data management plans help make clear plans for the study: *"I think this is important. You need to have a clear plan [on] what you're going to do with your data."* [I4]. However, researchers with more open data practices, such as [I8], [I9], and [I10], find the data management plans a burden, and state that the data management plans are not useful in practice. [I9] states *"I can't stand those things. […] I always ask my data manager to fill it out."* [I9]. [I8] states, *"I think that in theory is good, but in practice, it's a burden, […] because it doesn't prove useful, [it] has never proved to be useful. […] It's just a waste of time."* [I8]. Similarly, [I10] understands their purpose in theory, but they think once the plans are done once or twice, then the data management plans seem to turn into a bureaucratic burden: "*There are rather a burden I must say. […] I understand why they do it. But I think, once you've like done it once or twice […], I mean, you work along certain principles and the fact that you have to write this upfront… It [the plan] always changes throughout the course of your research because that's what happens, you change your plans slightly because whatever reason. So, it's another bureaucratic tiger, I would call it."* [I10]. Therefore [I10] questions whether writing everything upfront makes sense regarding the concept of data management plans. [I10] advises making data management plans shorter, at least for the beginning of the research, and asking the researcher to get back to it once the plan becomes relevant toward the end of the research. Because of mixed attitudes towards data management plans, our analysis only indicates a low influence of this instrument on open data practices.

4.7.2.2. Instruments that actively support researchers in sharing and using research data

Data steward support

Our analysis finds evidence that researchers would highly benefit from more engagement with data stewards for open data practices. Receiving support from data stewards is found to have a very high influence on open data sharing and reuse in our case.

Data stewards are found to benefit researchers in many ways. [I1] states that data stewards are there to answer questions in their organization, and that receiving support from data stewards is important because oftentimes there are problems with dealing with data that are hard to understand. [I10] brings attention to how the role functions as a point of referral when you need help: *"It's [data steward] more approachable, I would say, […] And then if they don't know [how to help], they can send you to someone else. As an entry point, it's useful."* [I10]. However, our first observation is that not every organization has a dedicated data steward role. This could also be the reason why some interviewees initially were not sure about the exact role of a data steward during the interviews. [I2] and [I3] state that in their organization, there is no formal data steward role. However, despite this, [I3] states that they have colleagues who have a lot of knowledge on data-related subjects and [I3] can easily go and ask questions to these colleagues. [I2] states that the policy in their organization is that instead of having a fixed data steward, everybody (individual researchers) is held responsible to uphold that stewardship. Whether the point of contact is called a data steward or not, it is apparent that the majority of the researchers express the need to be able to refer to a person of contact when they need help.

Some interviewees are not sure whether they have data stewards to who they can refer for data-related questions in their organization [I4, I5], which could suggest a problem concerning whether researchers have enough awareness of the existence of data stewards in their organizations. Regarding this, [I11] states that organizations should build more awareness of the fact that there is support available (about research data management topics). [I11] states that there is currently a *"lack of knowledge about the fact that there is support available*" [I11]. If more researchers know that they can indeed get support for problems they are facing, this could stimulate open data practices [I11].

There is again a pattern where researchers who have less experience with open data practices are less aware of the function of data stewardship (role) and also have trouble evaluating whether they would need help from such agents, especially regarding open data practices. The researchers who have more experience with open data practices have stronger opinions about the importance of the role of a data steward in the name of open data practices: [I9] explicitly calls the role of data stewardship essential for open data practices. [I8] states that they currently do not receive enough support from the data stewards in their organization, and that the support from the data stewards should be enhanced to reach better open data practices. [I7] gives a similar statement, stating that in their organization data stewards are very busy, and they do not have time to personally look at the data you have: "*I don't think it's [the support from data stewards] sufficient, […] because they're so busy, they don't have time to, personally, properly look at the data that you have […]. If you have a very straightforward data set, then that's okay, because […] they've done probably hundreds of cases, but the moment your data set is a bit more complicated, or there is anything that's not standard, I think they don't have enough time to properly support you in case you would want to do something, for example, anonymization."* [I7]. [I7] then gives an example of an incident where they wanted to anonymize a certain dataset to make it open, but they did not get enough support to deal with this procedure. It is noteworthy to mention that several researchers expect data stewards to work on specific datasets or individual projects in detail (e.g. for data anonymization), while

traditionally, data stewards do not take these roles in universities (formally). This expectation could be the result of not having (enough) data managers in these organizations. Regardless, our analysis shows that receiving support from data stewards has a high influence on open data practices.

Furthermore, it seems that there could also be a benefit in enhancing the role of the data stewards towards taking a more concrete (supporting) role in open data aspects. During the interview, [I2] shows the profile of people who take the supporting role on data management topics in their organization, and states that although there are a lot of people being responsible for data management topics, there is nobody who is explicitly tasked with "open data" aspects, which [I2] finds problematic in the name of open data practices. It could indeed be important to ensure that researchers are aware that there is somebody in the organization who can provide support on specific open data issues. However, since information on the exact role divisions in these organizations (where the interviewees work) is not available to us, we cannot draw conclusions on how to restructure/respecify role divisions.

Working with data managers

Being able to work with data managers to whom they can shift their research data management responsibilities seems to be a valuable instrument for researchers, especially those that engage with open data practices more than the others. Data managers are reported to be the primary agents that "look after" the datasets and keep them "up to date", which suggests that their role is vital in ensuring the data can be reusable for open data practices [I7]. However, many interviewees state that they cannot work with data managers [I1, I2, I4, I5, I7], because there are no financial resources available to them that would allow working with these people. [I7] states they have tried working with a data manager for their study before but could not do it because of financial issues. In a few departments, the departments themselves hire a data manager to work for the entire department by allocating their time to different projects [I2, I9].

Our analysis shows that mostly there is not enough budget for hiring data managers to get help for open research data practices. [I4] states that it is only possible to hire data managers if there is enough research money since researchers probably cannot get money from their department to work with data managers. [I2] notes that although there are a few data managers in their organization, these people only make sure that the collected data get into datasets, because there is not enough money to ask for more activities, especially open data activities: *"They [our data managers] only work on making sure that the data that we collect gets in the datasets. That's their level of activities because we can't pay them. […] But we don't have money to do the next step on 'opening up' the datasets, even if we wanted to."* [I2]. In the interviews, apart from having insufficient financial resources, no other reason is reported on why, currently, access to this instrument is low. [I2] also adds that the (open) research data management activities are not hard to learn, and the topics can be understood in a couple of courses. Therefore, apart from the money issues, hiring somebody for research data management is easy in terms of ensuring that the person who is hired has the right qualifications.

Researchers who are in more senior roles and who participate in larger studies seem to have more access to data managers. [I3], [I8] and [I9] express their satisfaction in being able to work with data managers for building the databases and other specialized data work (e.g. imports/exports of data). [I4] states that there is a culture in their organization that implies that you are somebody who "needs" a data manager (only) if you have a big project: *"Well, if you have research money [you can have data managers]. But I'm not sure if I could get money from my department to do that. […] I do see that there is a […] culture: […] if you have a big project, you need a data manager. […] But you need research money to do that"* [I4]. [I5] also confirms that only in big projects hiring a data manager may be a possibility, and adds that if more financial funds existed, this would positively affect open research data reuse, because data managers shape the data into standardized formats (which make them reusable). [I4] also states that a barrier to open research data practices is not having money to hire people for research data management activities. [I10] states that nowadays, if they write a grant application, they already reserve a budget for research data management. However, it seems that most grants do not allocate money for such activities. [I7] states that the reason why grants often do not include money for these data management activities is that data management is seen as a burden: "*All the way back, I think data management is something that people, still, view sometimes as a burden and something that you have to tick the box and then you can go on your merry way. And that's also why often in grants […] there's not money requested specifically for this kind of people to have them [data managers] on your project."* [I7]

Our analysis suggests that data managers do have the potential to positively affect open data sharing if financial situations were different. The instrument is found to have a very high influence on open data practices as many interviewees explain the importance of working with data managers on open data practices, despite not being able to access it.

Enhancing the library's role in support for open data practices

Several researchers state that they currently do not receive any help from the libraries in their organization [I3, I4, I8, I9], although some researchers say they have previously interacted with libraries regarding open data practices, such as [I7] who states they recall taking support from libraries regarding privacy-related issues in open data practices. [I7] also adds that the library in their organization provides courses on open data topics. [I8] states that they would appreciate getting more help on technical issues from the library regarding open data practices and they believe library support could be an important element for open data practices. On the other hand, some statements imply that there is no need for more library support. [I9] implies that especially regarding technical support, the role of the data management department in their organization fulfills these needs, and therefore there seems to be no reason why they would expect support from libraries on these aspects.

However, [I3] brings attention to the fact that the supporting role of agents like libraries could increase significantly in the future if open data practices get more widely adopted, because these practices require a lot of work and there is a lack of financial resources (such as grants): "*For the current situation [the level of open data sharing], this [technical support] is fine. When the world will change and [more] open data sharing should be there, […] then more*

support is needed. Because it's a lot of work and it's melting the grant I receive. There's not enough money to do that work." [I3]. This suggests the library's role as a supporting agent could be more pronounced in the future, especially for technical aspects.

For this specific instrument, we believe there needs to be more data to judge whether researchers would benefit from help from libraries on either technical or non-technical aspects of open data sharing, because during the interviews the interviewees did not explicitly talk about their specific expectations from the libraries, although we intentionally gave them specific examples on what technical support or nontechnical support could potentially be (see appendix B for such examples). However, considering that several interviewees mention the need for more ICT support, libraries could be important agents for taking such responsibilities.

Support on legal aspects of data sharing (e.g. on compliance with privacy regulations)

As we explain in chapters 4.6.3. and 4.6.4. the obligation to abide by privacy regulations (i.e. complying with the GDPR) is cited as a strong barrier to open research data sharing in the case. Several researchers state that their organization provides some level of support for understanding and fulfilling these legal obligations regarding openly sharing or reusing research data [I5, I7, I8, I9, I10]. For example, [I5] states that there is one person in their team actively checking compliance with privacy regulations. The responsibility of checking whether the GDPR is in compliance falls under the data privacy or data security officers in the organizations [I2, I9]; and researchers report having engaged with privacy officers when they needed support on legal aspects of data sharing.

We identify this instrument as one that has a high influence on open data practices in our case for two reasons. In our analysis of the influence of this instrument, our first finding is that researchers think inadequate assistance is given from their organizations. Due to resource (e.g. time) restrictions, in practice, it is questionable how much useful support the privacy officers currently give to the researchers. [I2] states, *"So in reality, it's difficult to get useful information from them [privacy officers] because there were only two or three of them. And they're overloaded with work"* [I2]. The second finding is regarding several researchers explicitly stating that their engagements with legal teams or privacy officers in their organization often result in negative outcomes (i.e. the data not being (openly) shared), and that they feel as if the legal teams are not really supportive towards (open) data sharing [I7, I8]. [I7] states that researchers in their organization may have a will to share their research data, but legal teams seem to always focus on the "negative" or focus on giving a "no", since they see always issues with privacy: *"It's just that they are very strict on the legal issues. They are the ones that are saying, 'you should have had informed consent from everyone before you can do anything with your data' […] They're trying to make sure that there is no liability at all, which I understand, but that also makes it very difficult as a researcher […]."* [I7]. [I8] states that the legal teams in their organization are "blocking" them from sharing their research data: *"They want me to adhere to legal guidelines, but at the same time they make it unbelievably difficult for me to do that"* [I8].

Our analysis suggests that researchers currently view legal teams, and data privacy/security officers as agents who always make it harder to (openly) share research data. This is also in line with [I11] stating that currently, the main goal of these agents (i.e. people who work on privacy aspects in research data management) is making the data comply with FAIR principles, and that openly sharing the data only comes as the secondary goal only when the privacy criteria are fully met. Therefore currently, there may be problems regarding the communication between researchers and legal teams.

Providing training and educational support

Many researchers state that they have received some sort of training or education from their organizations, mostly from the libraries, regarding open science and data management [I2, I4, I6, I7, I8, I9]. Such trainings are on privacy, data protection, and ethical concerns on data handling in clinical research [I4, I6, I9], courses on open science [I1], and data management [I2]. Our analysis indicates that this instrument has some level of influence on open data practices, mainly because some researchers state that they find these trainings beneficial for open data practices [I9, I8]. However, our analysis also suggests that currently the existing trainings are not always perceived as perfectly useful or purposeful, because several researchers chose to talk about their criticisms over the available trainings at their organizations and about how these trainings should be restructured or retargeted to achieve better effectiveness on open data practices.

Researchers express their criticisms over the trainings available at their organization and suggest ways of improvement: [I2] brings attention to the fact that their organization does not have a dedicated open science curriculum (i.e. there are elements of open science in other courses, but there is no uniquely identified course of open science) and also that the open science-related courses are currently not mandatory to the researchers. [I8] criticizes that educational efforts in their organization are too general. [I8] then advises that trainings on open science and data management should be tailor-made so that the knowledge obtained from these trainings can be applied to specific projects that researchers are conducting.

Open science trainings and modules in organizations aim to cause a motivational shift for researchers towards openly sharing their data. However, our analysis suggests a possible issue that the target group of these trainings may not hold the executive power to openly share research data. Researchers often indicate that the trainings in question are targeted at PhD students [I9, I7]. [I7] brings attention to the fact that the trainings are given to PhD students (or people starting their postdoc), even though these people do not have the executive power to make decisions about data sharing: "*When I was a PhD, there were PhD-specific trainings on, for example, open science. The difficulty is that they give these trainings to me as a PhD student or a starting postdoc. But I don't have the executive power to make the changes… So the open science [education] was […] 'you should always share your data and make that possible'. I'd love to… But if my supervisor says no, then the supervisor says no. […] So there's a lot of training, but it might be nice if you give training also to the higher-ups who can actually exert power to do this"* [I7]. This suggests the need to focus on giving trainings also to people in senior positions to stimulate a top-down motivational flow in organizations.

4.7.2.3. Instruments that relate to financial resources

Providing separate funds for treatment, management, and sharing of open research data

None of the researchers is aware of any separate funds for treatment and management of openly shared research data in their organization. Several researchers point out that nowadays grants sometimes include compensation for data management activities, although most of the time, especially regarding open data activities, this is not the case [I2, I3, I5, I7]. [I10] states that researchers who wish to have money for these activities should make sure to reserve these funds in the grant application.

In theory, covering the cost incurred during the process of preparing a dataset to be published openly would alleviate the burdens of open data sharing. Our analysis also shows evidence that having an access to such funds might translate to open data practices. [I4] and [I5] both state that the availability of funds would definitely change motivations towards open data sharing: after all, if you need to take money out of your own research money for something (i.e. open data sharing) that is not a requirement, then researchers will not feel motivated to make their research data open.

However, we recognize that our indication that this instrument (to some extent) influences open data practices is because some researchers state that more financing would motivate them towards open data sharing (i.e. building a causal relationship). This does not mean this relationship would be strong enough to make drastic changes in the actual open data practices. Regarding this, [I2] states that having separate funds for open data practices would help, however providing financial instruments would not "solve" the problem (i.e. lack of open data sharing) in the field completely because the dominant barrier to open data sharing is that most researchers do not believe in open data sharing as a necessity. Similarly, [I9] also states that they do not consider lack of financing to be the biggest bottleneck in front of open data practices in the field.

An important note for this instrument is that alleviating the monetary costs of open data practices may not necessarily mean alleviating the timewise cost of open data practices, especially in a field like Epidemiology where there is clinical work. Several participants reported "not having enough time for making datasets open" as a strong barrier to open research data adoption. Many Epidemiologists have extra time pressure due to having clinical work (which has a distinct feature of being a work type that cannot be delayed) on top of their research activities. As the issue of not having enough time may be more pronounced in this field, providing financial funds may not necessarily compensate for the extra time that a researcher has to allocate for open research data sharing practices [I2]. If lack of time is not something that can be addressed by funds, a more detailed investigation on how to address the time problem of researchers may be needed to enhance open data practices.

4.7.2.4. Instruments that build a culture of data sharing and create incentives

Recognizing and rewarding open research data sharing contributions

The majority of the interviewees state that data sharing contributions are not recognized or rewarded in the field of Epidemiology. [I5] states, "*I think there is no recognition at all. […] There's no recognition for anything but writing papers basically in my field."* [I5]. [I8] states that they wish there was more recognition for data sharing efforts since these activities make up a large part of their work. Several interviewees explain how more rewarding and recognition could lead to higher open data practices in the field. There is also evidence that the field may be, although slowly, evolving towards more recognition and rewards for such contributions. [I10] states, "*I would say that's [recognition and rewarding for (open) data contributions] what everyone wants it to be like. What is now we rewarded is publications in high-impact journals, and […] the number of citations, etc. And I think the field really wants to move towards [having] the number of data sets that you have provided for open access and the number of times those data have been reused as a sort of a metric. But it's not really there yet. It's moving slowly"* [I10].

In theory, a way to recognize and reward data sharing contributions is to provide and use metrics that incorporate data sharing contributions (as opposed to metrics such as journal impact factors that focus on just publications). In different fields, such as the fields of Ecology and Evolution, new author-level metrics that value the dataset output of the research have been proposed with the argument that better recognition via these metrics will incentivize data sharing (Hood & Sutherland, 2021). In our case, interviewees state that they are not aware of any track metrics that incorporate data sharing contributions in the field of Epidemiology [I1, I3, I5, I6, I7, I8, I9]. Therefore, it is not possible yet to understand whether such metrics could influence open data practices in this field.

However, one notable mention in our case is the Recognition and Rewards programs which are programs implemented by the universities and research institutions in the Netherlands to change the way researchers' work is evaluated in academia (Miedema, 2021). [I2] talks about how their organization is in the process of adapting to the Recognition and Rewards program, although there is still a way to go. Recognition and Rewards programs gained relevance in the Netherlands in the last years due to the argument that conventional rewarding mechanisms such as journal impact factors do not match with the open science paradigm (Miedema, 2021). For example, in 2021, Utrecht University shared the vision for its Recognition and Rewards Program. The university talks about possibilities for changing the assessment criteria and "moving away from simple counting, now requiring narratives and indications of societal impact" (*Open Science: Recognition and Rewards*, n.d.). The programs will give researchers a chance to provide narratives on the entirety of doing research, from defining questions to research output to research effects and education (Miedema, 2021). Since these programs have been very recently introduced, it is not yet clear if and how they could properly incorporate contributions of open research data sharing. Our analysis shows that the implementation of these programs should be observed in the following years.

Considering data sharing contributions during hiring, tenure, or promotion decisions

A specific instrument that intends to reward data sharing contributions is incorporating them in hiring tenure or promotion decisions. Utrecht University's webpage for its Recognition and Rewards program explicitly states that UMC Utrecht has already begun implementing the program for promotion and tenure systems (*Open Science: Recognition and Rewards*, n.d.). However, the interviewees from this organization did not mention this during our study. This could be due to the novelty of the program. Similar to our discussion in the previous subsection, it is also not clear yet whether these new systems properly incorporate open research data practices, and this needs further examination as the programs progress.

Nevertheless, the majority of the interviewees state that data sharing contributions are not considered during hiring, tenure, or promotion decisions in their field. [I1] states, "*I don't think they are priority now. I think […] Yeah, people would be keen on doing [open data sharing] if [this was a priority]. […] Because [currently] you are […] punished if you openly promote your data and don't get cited, for instance. […] In the end, researchers are very much judged on how much they are cited. […] So, if openly sharing data was part of the main criteria to promotion decisions, they would be encouraged to do that [open data sharing] and do it more often."* [I1]. The instrument is indicated as one that has medium influence in our case since there have been some statements describing how the availability of this instrument would incentivize open data practices.

Requests for open research data sharing from organizations, funders, journals

Almost all interviewees reported having been asked to openly share their research data either by the funder or the journal they worked with before. For example, The Dutch Research Council (NWO) asks researchers to share their data openly as much as possible as part of its guideline that came into effect 5 years ago (*Other Researchers Can Use Your Data without Having to Repeat the Animal Testing*, 2021). It seems that research organizations (where interviewees work) do not request open research data sharing from the researchers. [I10] states, *"[whether they request for it or not] depends on the journal. Some of the journals, they really require it nowadays, at least the metadata."* [I10]. The common practice for journals in the field of Epidemiology is requesting the researcher to make the research data open, and if this is not possible, provide an explanation why that is not possible: *"They [journals] stimulate this, and they at least require you to provide a statement whether it's available and if not what, why it isn't available"* [I1]. Therefore, the requests for data sharing from journals are not mandatory. Regarding grant providers, it seems that grant providers are more encouraging than journals: *"They [funders] do [ask for open data sharing]. [For them] it's really important*." [I9].

Regarding how influential these requests are on the actual open data practices, we found evidence that a researcher may decide to share their research data upon being asked by the journal: [I9] states that they have had this experience, they once made a dataset open solely because the journal asked for it. However, requests may also get shadowed by other stronger factors influencing motivations of open research data practices: For example, [I3] states, *"Some journals do ask [for data sharing] or statements about, like the BMJ for instance, and other*

journals. […] Usually, I say then that the data are available on request. […] sometimes I have to add an extra sentence about privacy issues of patients- that you can identify patients [if data were made open]. So, therefore, I'm not allowed to put it completely out. […] It's okay for a journal to ask for it. It doesn't change my behavior." [I3]. Since the majority of the interviewees do not talk about how these requests relate to open data practices (in practice), the instrument is marked as one that has a low influence on open data practices in our case.

Demonstration of benefits of data sharing and the need for data sharing

Researchers need to see the benefits of open data sharing to build motivation for the actual practice. In our case, we observe that many researchers do not think their organization helps them to comprehend these benefits and needs [I2, I5, I7, I8, I9]. There is also evidence that researchers have expectations from their organizations regarding this topic. [I2] states, "*There is minimal effort on those open science aspects. There is a lot of data management aspects on getting that right, but open [science] aspects […] are only minimally present [at the organization]."* [I2]. [I4] states that their organization does not sufficiently help them understand the benefits, and states, "*I see that people talking about it now. […] But it's not implemented truly yet. […] I think this culture needs to change and people need to really understand the benefits of openly sharing [data]*" [I4].

It seems that the organizations aim to help researchers in comprehending the benefits of open data sharing via courses or workshops. However, it is doubtful how useful these courses are in practice because courses may have limited reach to researchers: "*If you do these courses, […] where they tell you about the benefits and why would you do it… But of course, if you're not interested, then you're not gonna do the courses. So you're not gonna know." [I7].* [I9] states, *"Maybe they [courses or seminars] are there, but I don't go to them*" [I9]. Our analysis indicates that different channels of communication should be considered to have better access to researchers.

Clarifying the concept of data ownership

In the Netherlands, the common practice of university policies regarding data ownership is that the university is the owner of the research data that are collected by any employee, which includes the researcher [I11]. For example, the Utrecht University's webpage for research data management states, *"Officially Utrecht University, as your employer, is considered the rights holder to the research data you create. You, as a researcher, have the primary responsibility for taking care of the data"* (*Research Data Management Support: Policies, Codes of Conduct and Laws*, n.d.). However, an issue that is brought up in the interviews is that many researchers are struggling with this and many researchers believe that they "own" the research data. Our analysis indicates universities may be responsible for clarifying this to the researchers: [I5] states that "*if there would have been [a way of] making senior researchers […] change their perspective on what the ownership of data entails"* [I5] then this would have a positive influence on open data practices. One way to provide this clarification is by giving training modules: [I4] mentions that there are trainings given by human resources in their institution regarding this, and having taken these trainings, the concepts are clearer to [I4].

On the other hand, the responsibility of the research institution on explaining the concept of data ownership to the researcher could be tougher to fulfill if one argues that the current definitions and conceptualizations that organizations give in their policies may not be in line with the open science paradigm in the first place. [I11] states, "*I think we do struggle with […] just defining what data ownership is within the university. […] I do know that we [the organization] do need to define that better. Even within the university, most researchers think of the data that they collected, they being the owners of it, when in fact it's actually the university who by default is the owner of it. […] So then what it means to own something […] seems to have either very little meaning […] when there's a patent or an intellectual property to be taken care of […]. I think we need to think away from ownership. Especially if as universities we want to be able to [do] open science, it makes little sense to say then that 'we are the owners of the data' if they were promoting researchers to make this data openly available […]. What does this ownership actually mean and why are we so fond of it? […] It needs to be a lot made a lot clearer to researchers and even to the university boards. […] Even within the GDPR, for example, there's no such thing as ownership. It's about controllership. So maybe we should look at it in the same […]"* [I11]. This suggests that policies may need restructuring towards better definitions of what it means to "have", "create" and "control" data to increase open data motivations of researchers.

4.7.2.5. New instruments that emerged through the case study

Building an open science community within the Epidemiology field

Relating to building a culture of data sharing, a new instrument emerged during the interviews, which is building stronger open science communities, particularly within the field. Several interviewees express their expectations from open science groups, and state that if open science communities were stronger in their field, there would be more tendency for open data practices. [I2] states that open science groups have the potential for bringing attention to the theoretical and moral grounds that support open data practices. However [I2] also mentions that currently there is a lack of impact of open science groups on the field, because, for instance, there is no "open science officer" and no "open science group" in the department, and because the open science community in the institute only has a few active members; all of which are reasons for having limited open science culture. [I1] mentions that open science groups, although limited in number, do have an influence on other researchers in practice. [I3] states that the open science working group in their department engages with other researchers and that this group sets the stage for valuable discussions concerning open data practices. [I5] states that there was a very small group of people working in open science in their department, but this group was too small to make an impact.

Increase communication among the scientific community so that researchers with similar research interests/outputs are aware of each other

[I7] states that a barrier is a lack of communication in the scientific community, because researchers do not have awareness of whether other researchers are producing research that is similar to theirs. In that regard, an institution tool would be increasing awareness of which data researchers are working on, although how this can be achieved is not yet clear. [I7] explains

the issue by stating *"[for example] UK Biobank does this: if people submit the same proposal or similar proposals, they link up to two people and say 'you're actually doing the same thing or a similar thing. Can you work it out together so that you don't do the same thing twice?' […] A barrier [is] where we don't really communicate what we're doing always, and also people that write out grants might not always check that. […] If you don't know that other people are actually doing something similar, that's why you start collecting your own data instead of reusing [somebody else's]." [I7].* Our analysis suggests looking for methods to facilitate this certain kind of earlier communication in the scientific community, so that researchers can have a chance to know what their peers are working on (long before the release of associated journal articles).

4.7.3. Barriers to open research data sharing and reuse in Epidemiology

During the interviews, we asked the participants about the leading barrier(s) in front of open research data sharing and reuse in the field of Epidemiology. While some gave a single answer, others gave multiple answers to these questions. Below the leading barriers to open research data sharing (Table 11) and the leading barriers to open research data reuse (Table 12) as identified through the case are presented.

Table 11 Leading barriers to open research data sharing as identified through the case (ordered from the most frequently mentioned to the least)

Barrier to open research data sharing	**# of respondents mentioning the barrier**	**Explanation**
Privacy Regulations (Legislation)	5 respondents	Having to comply with the GDPR prevents researchers from openly sharing research data (see chapter 4.6.3 and chapter 4.6.4.)
Perception of lack of time	5 respondents	Researchers report that they do not have sufficient time to allocate for open data sharing, such as the time to prepare datasets or maintain them on open data infrastructures. [I4] states, *"I can't imagine [how] a PI (principal investigator) having to -apart of all the tasks that he or she has- also would dedicate extra time to create this infrastructure. That is something that is time-consuming"* [I4].
Perception of data ownership	5 respondents	This refers to the belief of "I put the effort to collect the data, so it's my own data ". [I4] states, "*PIs (principal investigators) really feel that they own the data, because they got […] the grant money to do the research, they went to all that trouble to get this research moving, collecting the data from all these people. […] Then now they would just open it […]. [PI's think], 'what about all the trouble that I had to actually collect this data?'"* [I4].
Lack of acknowledgment and reward	4 respondents	Because there is no acknowledgment or reward system for open research data contributions, researchers do not get incentivized for open research data sharing. [I8] states, *"Nobody is eager to just give away the data without any condition. So there are conditions […] conditions can be coauthorship or money."* [I8].
The desire to publish results before releasing data	3 respondents	Researchers would like to publish the results of the research before publishing the data because of the concern that somebody may formulate the same research question and do the same research. This barrier is pronounced in Epidemiology as researchers may have flexible research agendas (i.e. new

		research questions occur during the process) and long studies. [I3] states, *"My studies are usually large […] I cannot specify all the research questions for myself before. My research agenda […] [is] flexible and it develops over time, and if my data are already out there, then I don't know what other people do with it, and then you get double work."* [I3].
Lack of financial resources	3 respondents	Researchers may not have access to financial resources (to hire people who can take care of data management activities) that are necessary to make a dataset open. [I4] states, *"There needs to be money to hire people specifically work with that [open research data management]"* [I4].
Concerns about the quality of the paper that will reuse the dataset (e.g. Flawed interpretation)	2 respondents	Researchers are concerned about the possibility that the quality of the papers that reuse their dataset will be bad, which is why they are not willing to put their dataset in to open domain. [I9] states, "*We made [a dataset] publicly available. Because we felt that it was the right thing to do. A year later, a paper was published with this data [which completely misunderstood the dataset]. So, you took my data. You wrote a terrible paper because you completely misinterpreted the data. […] And in those kinds of experiences, then you think 'I don't want to share my data because people are gonna do just junk with it'"* [I9].
Not having an open science culture	1 respondent	Researchers state that lack of an open science culture in the organization results in low open data sharing. [I2] states, *"[in my organization] there's no open science officer. There's no open science group. There is no open science interest group, the open science community […] has two active members […]. There is a limited open science culture*." [I2].
Satisfying the expectations of funders	1 respondent	Researchers feel that they have a responsibility to the funders regarding their research outputs, and making datasets open could jeopardize satisfying these expectations: *"[…] I have grants from the [a specific foundation], and I promise to do some work. And if I start collecting data, and if I put the data online, then other people can answer my question already. And I cannot comply anymore with my applications for the requirements of the […] foundation, so I cannot do it."* [I3].
Concern that there will be errors in the dataset	1 respondent	Researchers feel that it takes a lot of time till they realize that there is an error in the database. Thus, they are not willing to share the data before they work with the data in depth. [I3] states, "*Sometimes you are writing your paper no. 4 [4th paper], then you discover there's an error in part of the database which you didn't work on previously. So you discovered it [the error] only then. So you cannot share it [the data] before you really worked with all that kind of data. And if you share it, other people could get the wrong data actually."* [I3].
Not receiving research data management support	1 respondent	Researchers feel that they need support to deal with the management of open research data, and the lack of supporting agents such as data managers prevents researchers from practicing open research data sharing. [I5] states that a barrier to open research data sharing is*, "not having a data manager [that] you can easily contact with any questions or can easily help you […]. I think that also inhibits sharing data because […] data is somewhere on someone's computer, basically, not in a shareable format"* [I5].
	1 respondent	Researchers have to deal with the competitive atmosphere in academia. Competition is caused by the lack of funding that is available for research, which is why researchers have to "fight" for a place in academia. Therefore, open research data sharing is not a priority for researchers in this atmosphere. [I1] states, "*I found out […] gradually during my PhD is that it's [the academia] really competitive and that's because […] funding is an issue and that there is simply […] lack of money that is available for researchers to use*

Competition in academia and prioritizing publications		*and to get contracts for indefinite periods fosters this a little bit. […] Chances of getting this money [funding] are usually quite low […] Often people's […] positions really depend on this. […]. And if you don't get the [funding] money, you probably have to leave at the end of the contract […] This really sort of also stimulates people to boost their careers as much as possible on an individual level. […] You probably have to prioritize your own career."* [I1].

Table 12 Leading barriers to open research data reuse as identified through the case (ordered from the most frequently mentioned to the least)

Barrier to open research data reuse	**# of respondents mentioning the barrier**	**Explanation**
Inability to discern dataset content and data collection (data characteristics)	5 respondents	Researchers want to know exact details about the dataset that they will reuse. Researchers need a vast amount of detail about the data in Epidemiological research. If they cannot get full information about how certain measurements were made, then they are not willing to work with the dataset. [I9] states, *"If somebody says 'this weight of a person', was it measured, was it self-reported? If it's not mentioned, then how am I supposed to know how it was measured? How did they assess the weight? That's the problem with data dictionaries, they don't [do that]."* [I9].
There seem to be no (relevant) datasets available for use	2 respondents	Researchers cannot find (relevant) datasets available for reuse in the open domain. This barrier is enhanced by the fact that in Epidemiological research (such as in the field of infectious diseases), data easily lose value (relevancy) as time passes. [I10] states, *"I think that's what happened with COVID [research] a lot. People were searching for data everywhere, but they were just simply not there. Then by the time they become available, they're no longer relevant."* [I10].
Not having an open science culture	1 respondent	*See the barrier explained in the previous table*
Not knowing the source of the data	1 respondent	Where the data come from matters to the researchers. If they have certain doubts and concerns over the source of data, researchers may be less willing to reuse the dataset. [I3] states, "*if it's [the data] somewhere from someone I don't know in a strange country and whatever, I'm very unlikely that I will start using that [data]."* [I3].
Lack of awareness o about available (open) repositories	1 respondent	Researchers may not be aware of the repositories that are available for use and specific repositories that more relate to their research interests. [I7] states that the main barrier to reusing open research data is the *"[lack of] awareness about […] what are these repositories and where are the ones that are interesting for me"* [I7].
Issues with findability of the data	1 respondent	Researchers may not be able to find the data they are looking for since current search engines may not satisfy data search needs properly. [I10] states that the barrier to open research data reuse is "*to be able to find them [the data] […] [in these] data platforms"* [I10].

4.7.4. Comparing institutional and infrastructural instruments and discussing their interrelation

Our analysis suggests that in the field of Epidemiology, institutional instruments may have more influence on open research data practices than the infrastructural instruments. Reviewing the leading barriers in front of open research data adoption in chapter 4.7.3., it can be seen that the most frequently mentioned barriers are mostly those that relate to institutional instruments. For example, 'perception of lack of time' can be addressed by providing the possibility to work with research data managers, and 'perception of data ownership' can be addressed by organizational efforts on clarifying the ownership concept to the researchers. Similarly, lack of acknowledgment and reward, lack of financial resources, and not having an open science culture all relate to the institutional instruments that are evaluated as part of this case study. More importantly, the majority of the leading barriers indicated in our analysis contribute to the lack of a data sharing culture in Epidemiology, which we discussed in chapter 4.6.5.

Our case study analysis indicates that there are three essential barriers to data sharing practices in Epidemiology: the privacy regulations, the perception of data ownership, and the lack of an open data sharing culture. We argue that for now, focusing on the latter, building an open data sharing culture, is the most important step for the field of Epidemiology since this is considered to be the factor that is easier to tackle than the other. This deduction to prioritize focusing on instruments that enhance open data sharing culture also echoes through the statements of the participants: [I4] summarizes the issue the best: *"First of all the culture needs to change. So there needs to be some effort from the institution to talk about the benefits and why [open data sharing] is important, and that you are not at all the owner of your data. So bringing more awareness about it and that it is just fair that other people can use it. This is 'increasing awareness'. […] Then, after that, giving […] financial support and also infrastructure support, that's the next step. I think we are still in the first phase that you need to increase awareness. And then when people buy the idea, then [..] all the infrastructure and the support need to be there. […] In the field of Epidemiology, […] it's a starting process"* [I4].

The infrastructural and institutional instruments that are evaluated in this case study also relate to one another. For example, the instrument of providing education and training complements the instrument of providing metadata on data collection on the data platforms. Researchers should be given trainings on how to prepare data dictionaries that satisfy the needs of the field, and with these trainings they can then meet the data infrastructures' expectations from them. Similarly, the instrument of enhancing the integration between data infrastructures will require more engagements with data stewards and data managers, as these integrations are expected to result in researchers needing more ICT skills and knowledge in the research data management cycle. Combining instruments could then increase their effect on open research data practices since the system where open research data adoption occurs is complex, meaning that many factors (e.g. barriers) are interrelated and they have interactions with one another. For example, if (in the future) data infrastructures that provide a workaround to privacy concerns are developed and adopted more in the Netherlands, this will bring a new set of responsibilities to legal teams and data stewards in universities in terms of properly guiding researchers in using

these systems for sharing and reusing research data. Our case analysis also indicates that there is not a "one size fits all" solution to the problem. Policymakers should ensure that a variety of instruments that target different barriers (which we discussed in chapter 4.7.3.) are used to increase the open research data sharing and reuse levels in Epidemiology. Tackling only one of the barriers will likely not prove useful since the low level of open research data adoption is a multifaceted problem. Finally, it is also important to mention that the influence of the instruments depends on the current level of data sharing in the field. If this level changes in the future, instruments that are previously considered to have a low impact may gain more value and turn into more important tools for open data practices.

4.7.5. The conceptual framework refined by the case study

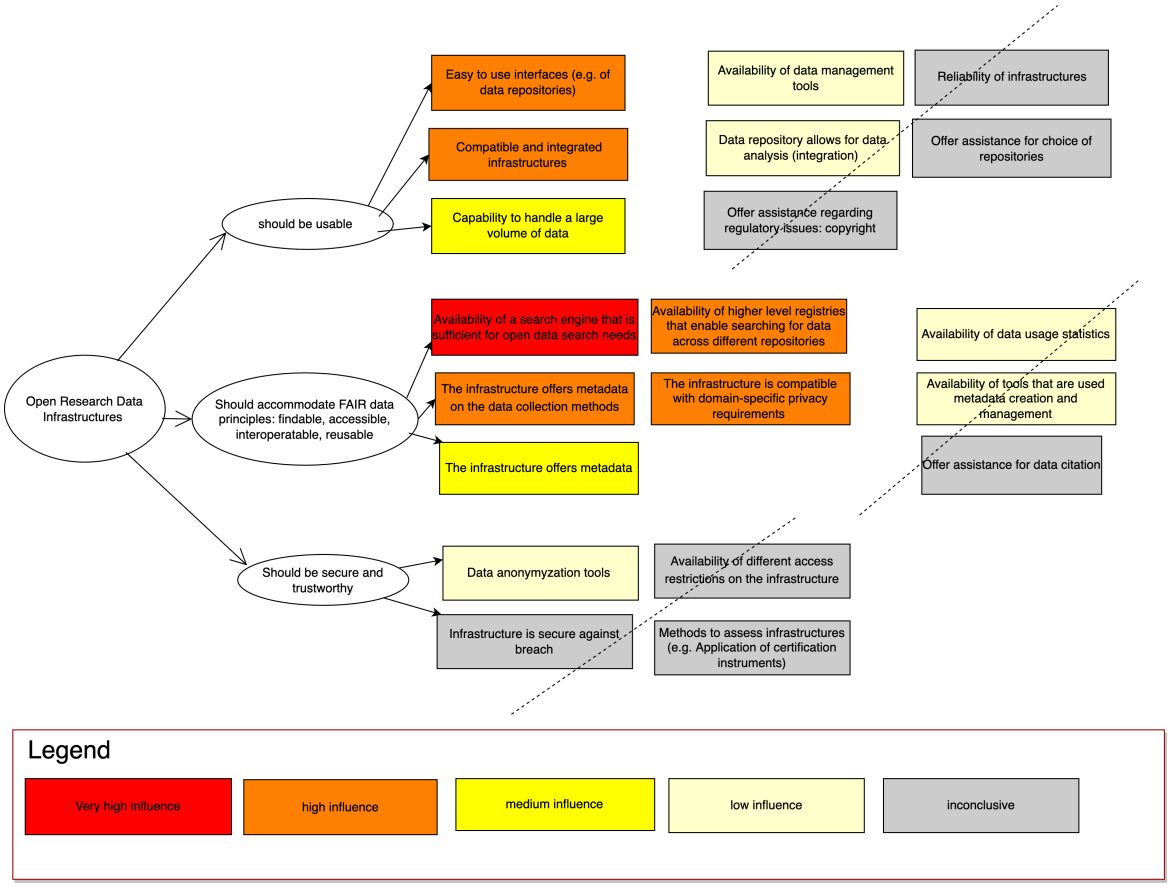

Figure 11 Refined conceptual framework for infrastructural instruments

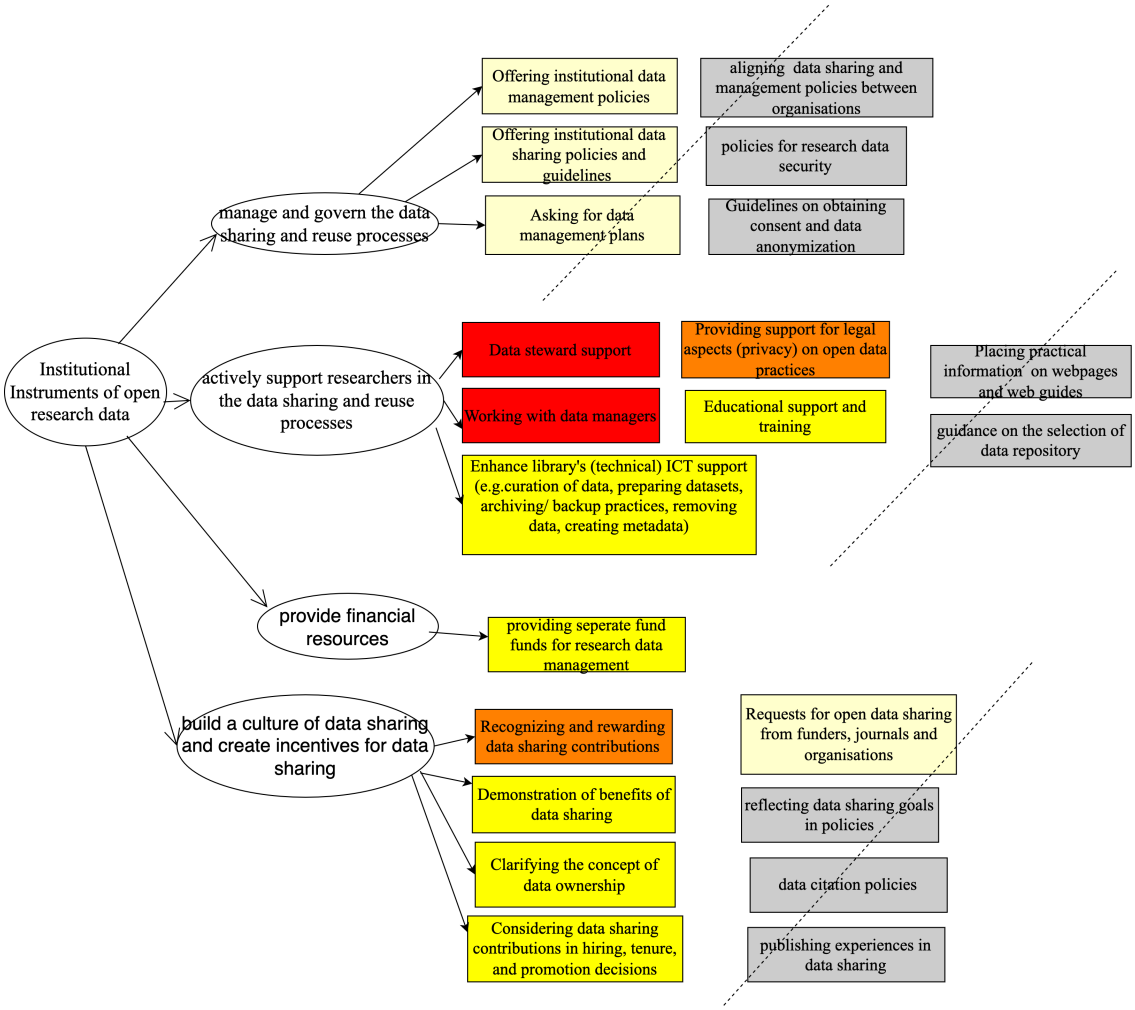

Figure 12 Refined conceptual framework for institutional instruments

4.7.6. Reflecting back on the theories

In our case study analysis, many of the results concerning how instruments affect motivations and behaviors in open research data sharing and reuse can be traced back to the theories we previously examined in chapter 3.1. of this report. For example, the importance that researchers attain to instruments such as offering metadata and metadata on data collection relates to the "Output Quality" that influences the "Perceived Usefulness" in TAM 2 model by Venkatesh & Davis (2000). Supporting the propositions of TAM 2, our analysis shows that the adoption of open research data in Epidemiology depends highly on the quality of the research data that can be retrieved from open data infrastructures, and we also establish that this output quality can be enhanced by providing detailed data dictionaries that describe the details of the data and the data collection. In Figure 13 below, we establish how some of the major results of our case study support the propositions of the TAM 2 model.

However, our analysis shows us the shortcomings of this model, at least in the context of our study. First, we realize that for open data infrastructures, the motivations for using the system "Intention to Use" cannot solely be related to "Perceived Usefulness" and "Perceived Ease of Use". In our case study, we discovered that several other important factors, such as one's

(perception of) lack of time to use the system (despite perceiving it as useful and easy), have consequences on the usage behavior, and this relationship cannot be captured through the TAM 2 model. Moreover, the model fails to incorporate the external conditions that limit usage behavior specifically by imposing rules on the system and the user. We observe many of the researchers use (or do not use) data infrastructures (e.g. certain online workspaces, data repositories, or analysis tools) due to established processes that are governed via the rules in their organizations. Therefore, we argue that the technology acceptance models, at least in our case study, fail to incorporate the (coercive) institutions which have consequences on one's act of participating in the system and one's acceptance of the system.

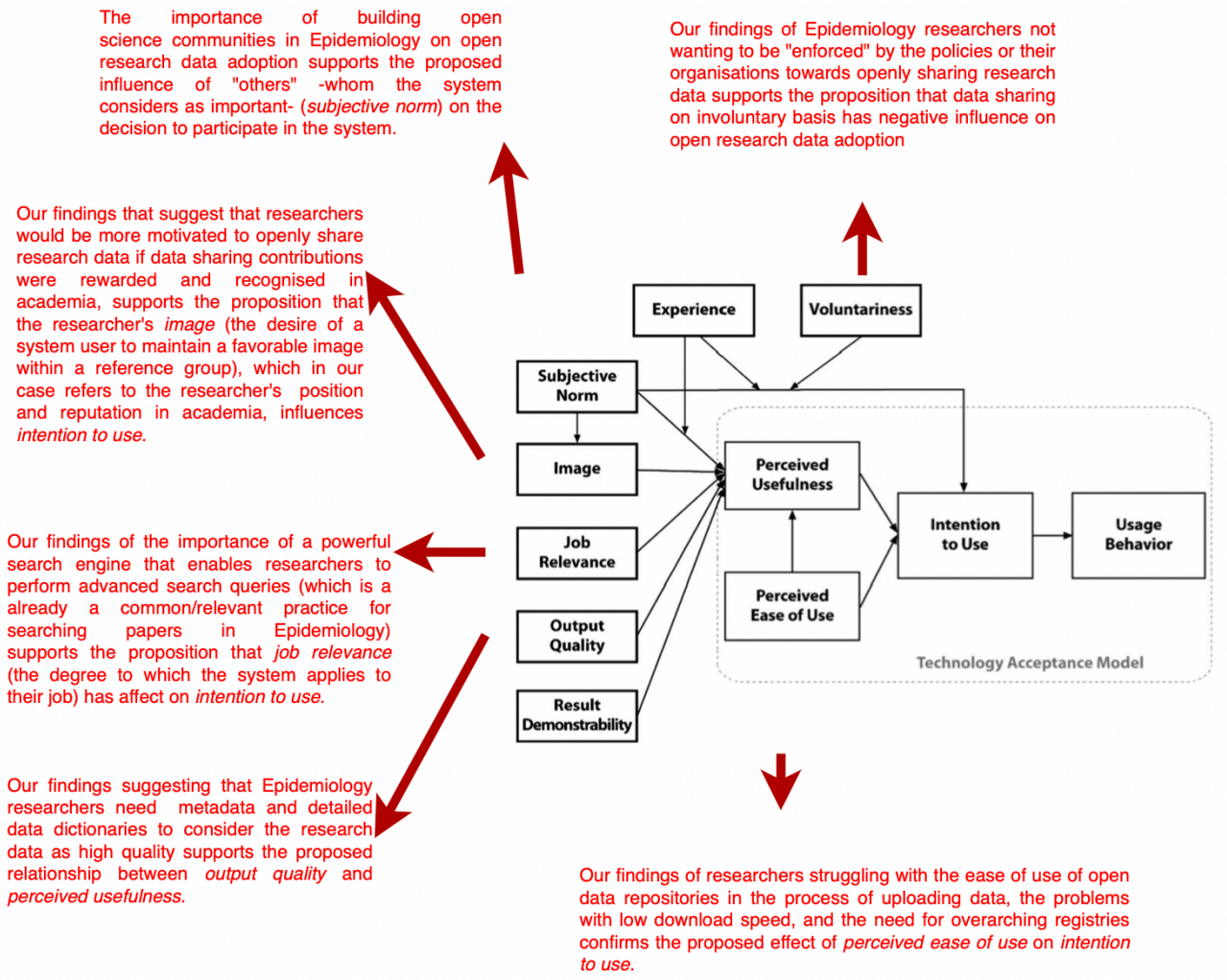

Figure 13 Reflection on TAM 2 Model

Furthermore, some of our results could also be conceptualized based on the institutional theory. First, we can conceptualize some of our findings through institutional pressures. With our case study, we establish that the introduction of the GDPR has put serious restrictions on the way data sharing occurs in the fields of public health. In the context of institutional pressures, this influence can be contextualized as a *coercive pressure*, which happens through political actors, which in our case is the EU lawmakers. This coercive pressure could be conceptualized as a force that lowers (open) data sharing and reuse practices in not just certain organizations, but

all the research organizations operating in the field of Epidemiology (or any research field that deals with personal and sensitive data on high levels). On the other hand, organizations themselves try to increase open research data sharing and reuse. Therefore, our findings imply the existence of various organizational efforts that aim to reverse the negative influences (such as the GDPR). Research organizations' embracement of open science principles in their policymaking, offering of supporting agents such as data stewards and data protection officers, recognizing and rewarding data sharing contributions, streamlining the overall process of research data management, and changing data sharing culture via explaining the benefits of data sharing are all organization-wide efforts to lead to a *normative isomorphic change* which essentially opposes the coercive pressures coming from the legislation. Therefore, our analysis illustrates the multi-actor environment of open data sharing and reuse (or in other words "the problem arena"), as one that attracts opposing influences from *coercive pressures* and *normative pressures* that bring change in organizations. We establish this multi-actor conceptualization in Figure 14.

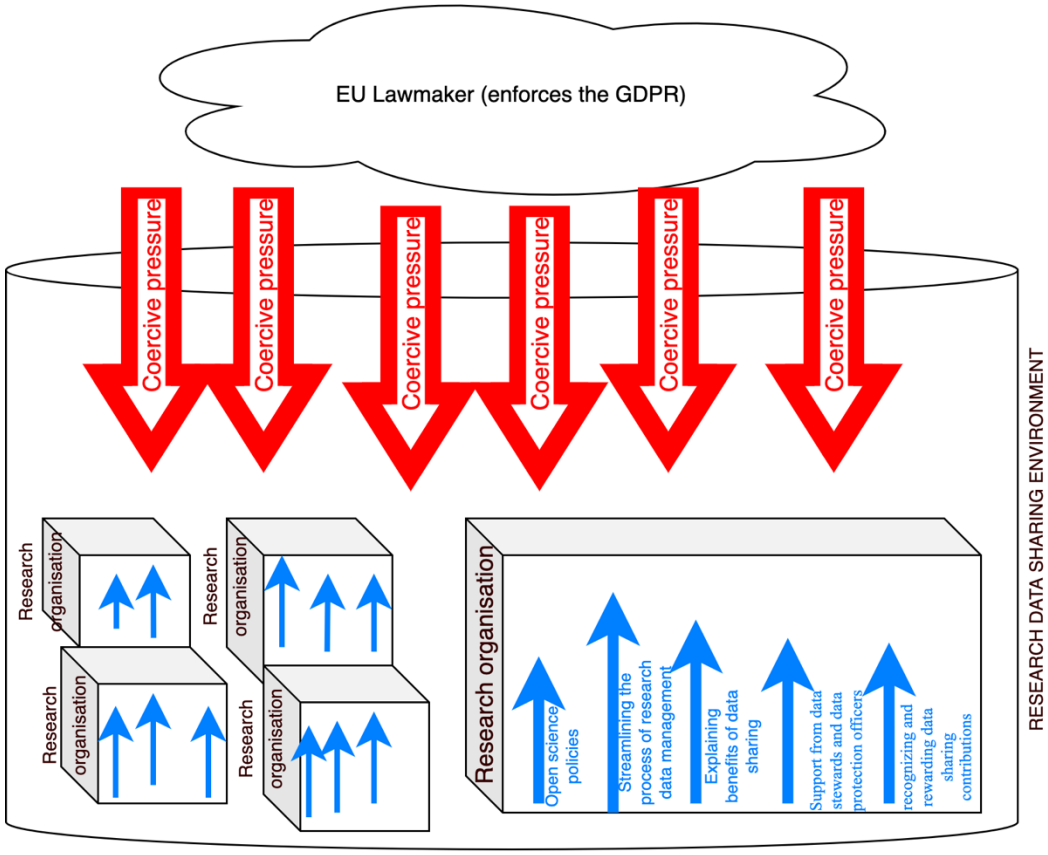

Figure 14 Reflection on the theory of institutional pressures (red arrows indicate coercive pressures that arise from legislation on a higher level, blue arrows indicate opposing normative pressures that arise from organizations)

4.7.7. Discussing what makes the case study findings specific/typical for the field of Epidemiology

Since our case study focuses specifically on the field of Epidemiology, in this section, we summarize how certain findings could be specific to the field of Epidemiology or the fields of public health.

1. The fields of public health are strongly bounded by GDPR in (open) data sharing because they deal with a lot of personal and highly sensitive data. This has consequences on the influence of support for legal aspects of data sharing as an instrument.
2. For Epidemiology, full anonymization of datasets lowers the scientific value of data significantly. Data linkages are important in maximizing the value of datasets in this field, but linking datasets when they are fully anonymized is a challenge. This has consequences on the influence of data anonymization as an instrument.
3. The fields where researchers have clinical work contain extra time pressure on the researcher. This has consequences on the value of instruments that provide financial resources (which would allow researchers to work e.g. with research data managers).
4. In fields where data collection may happen in clinical contexts, obtaining informed consent for (open) data sharing is hard: in clinics, it is perceived that there is often not enough time to attain a detailed informed consent process for every patient since hospitals, clinics, and overall, the healthcare sector in the Netherlands has been dealing with shortages of medical personnel for a long time ("Waiting Lists Still Increasing at Hospitals, Clinics," 2020). This has consequences on the influence of support for legal aspects of data sharing as an instrument and on the overall levels of data sharing practices.
5. For Epidemiology, understanding data collection methods is very important to researchers, and even very small nuances in the methods are considered to be important to know for the researcher. This has consequences on the value of metadata and data dictionaries as instruments.
6. Researchers in Epidemiology are very accustomed to making use of advanced search queries on search engines like PubMed when searching for references. These searches actually can be so advanced that there is a separate line of literature supporting researchers with search strategies (see Fatehi et al. (2014) and Motschall & Falck-Ytter (2005) for examples.) This has consequences on the expectations and value attached to the instruments that relate to providing powerful search engines for research data in the case.
7. Epidemiological cohort studies may take years, if not decades. The immense amount of effort that researchers have to put into collecting these datasets (both financial and time-wise efforts) enhances the beliefs of data ownership, compared to fields where data are collected from more centralized, automated sources, such as in the field of Astrophysics (Zuiderwijk & Spiers, 2019). This has consequences on the value of the instrument 'clarifying the data ownership concept'.
8. In Epidemiology, research agendas are likely to be flexible: researchers often develop their research questions at various stages of the data collection process. This has consequences on the low willingness to share datasets since there is always the possibility to work on the dataset more.
9. Epidemiology examines the health-related states and events that affect populations, such as pandemic and epidemic-prone diseases. These diseases sometimes emerge suddenly, such as in the case of COVID-19. Researchers in the field could often race with time when these diseases happen, which means the data sharing has to happen faster compared to other fields where human engagement is lower. Data lose relevance quickly. This has consequences on the level of support (legal, ICT, etc.) researchers need for research data management.

5. Establishing transferability of the case study findings

This chapter presents the workshop that is held to evaluate the case study findings in the context of establishing the transferability of the findings to other research fields, and validating our earlier deduction (in chapter 4.7.4.) about prioritizing institutional instruments over infrastructural instruments. The chapter respectively explains the motivation of the workshop, the background of the participants, the organization of the workshop, and the findings of the workshop.

5.1. The motivation for the workshop

We acknowledge that certain responses that we received in our case study could have also been provided for other fields because many of the barriers in front of open research data adoption also exist in other fields (Zuiderwijk et al., 2020; Zuiderwijk & Spiers, 2019). Acknowledging this, we propose that examining the extent to which the findings of our case study apply to other research fields is valuable. In line with this, to evaluate the case study findings, on the 1st of June, 2022, an online 1-hour interactive workshop was held with nine participants who either work as a data steward or a research data officer in a Dutch research university. As these participants are professionals that work in different research fields, the workshop intended to get their insights into the extent to which the influential instruments in our case would be useful in other research disciplines. This would enable us to discuss the transferability of the case study results to other contexts (i.e. research disciplines or the overall scientific community). Furthermore, the workshop also intended to get the participants' insights on which instruments would be more important than others, and why.

5.2. Background of the participants

We recruited participants that work in a range of different research fields. The participants of this workshop had backgrounds in a variety of disciplines: mechanical engineering, information management, microbiology, software engineering, genetics, and computational physics. Currently, these professionals work with different research disciplines in the institution: electrical engineering, computer science, quantum computing, mathematics, aerospace engineering, civil engineering, geosciences, policy and management, and mechanical and materials engineering.

5.3. The organization of the workshop

During the workshop, first, a 20-minute presentation was given on the research objective, research approach, and research findings (i.e. infrastructural and institutional instruments that are found to be important for open research data sharing and reuse practices in our case study). The rest of the workshop followed three activities that got insights from the participants. With permission from the participants, the workshop was recorded to later on make notes.

The first activity (Activity 1) was a 15-minute interactive activity that was held on a Miro Board. We asked the participants, "*Which infrastructural instruments could be also relevant for open research data sharing and reuse in other fields and why?",* and asked them to individually provide their insights using stickers that they could place next to the instruments they wanted to talk about (Figure 15). After five minutes of individual brainstorming, a group discussion was held where the participants shared their input with the rest of the group.

Similarly, the second activity (Activity 2) was also a 15-minute interactive activity that was held on a Miro Board. We asked the participants, "*Which institutional instruments could be also relevant for open research data sharing and reuse in other fields and why?"*, and asked them to provide their insights using stickers (Figure 16). Then, similar to the previous activity, a group discussion was held.

Finally, before the workshop ended, the participants were asked to fill out a short survey (Activity 3) asking them (1) *"What are the three most important instruments that enhance (or could enhance) open research data sharing and reuse behavior?"* and (2) *"Could you briefly explain why you chose these instruments, in 1-3 sentences?"* to obtain their insights on the importance of the instruments with respect to one another. The survey asked all the nine participants to individually choose the three most important instruments by giving them a list of all the instruments that we found to be important in our case study. We shuffled all the instruments to prevent bias when making a selection (Figure 17).

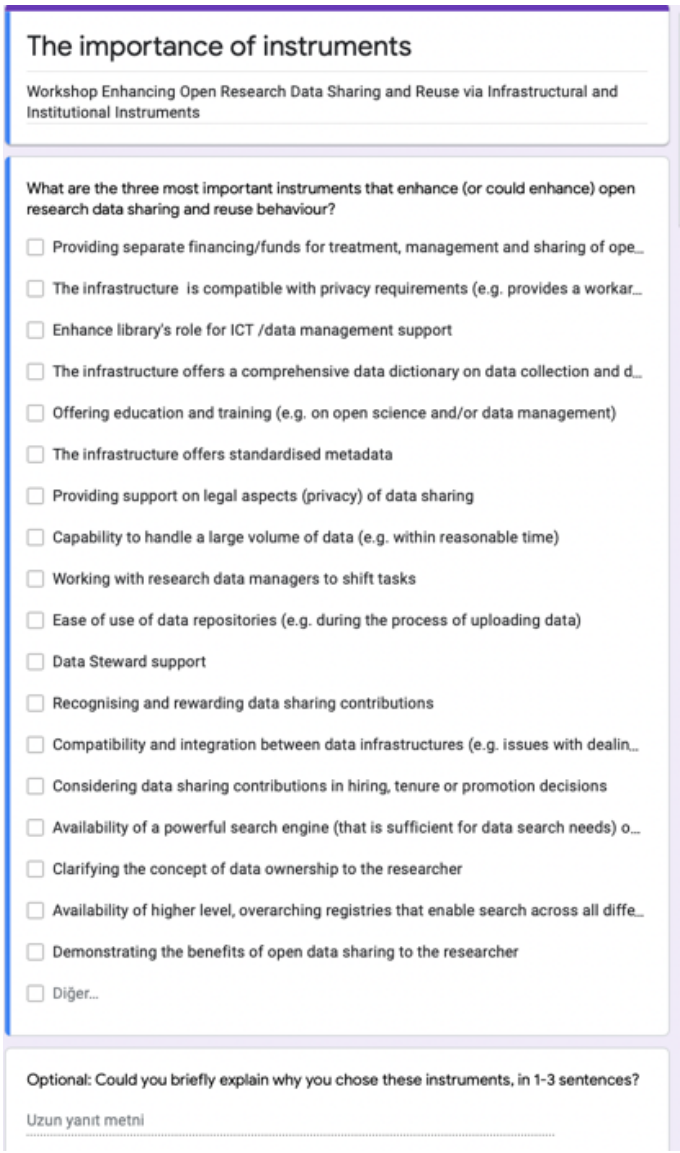

Figure 15 The survey questions

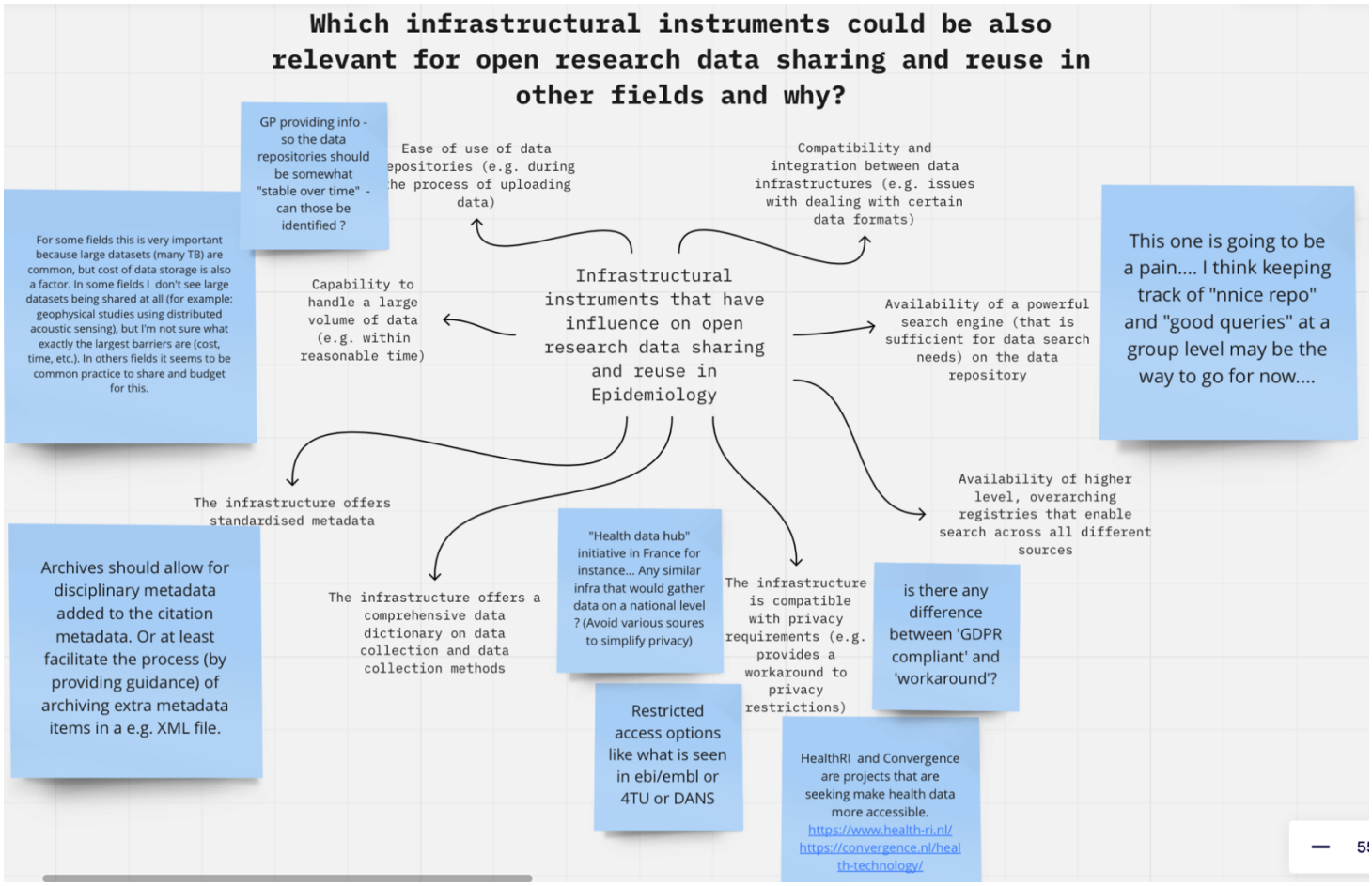

Figure 17 First activity: gathering insights on infrastructural instruments

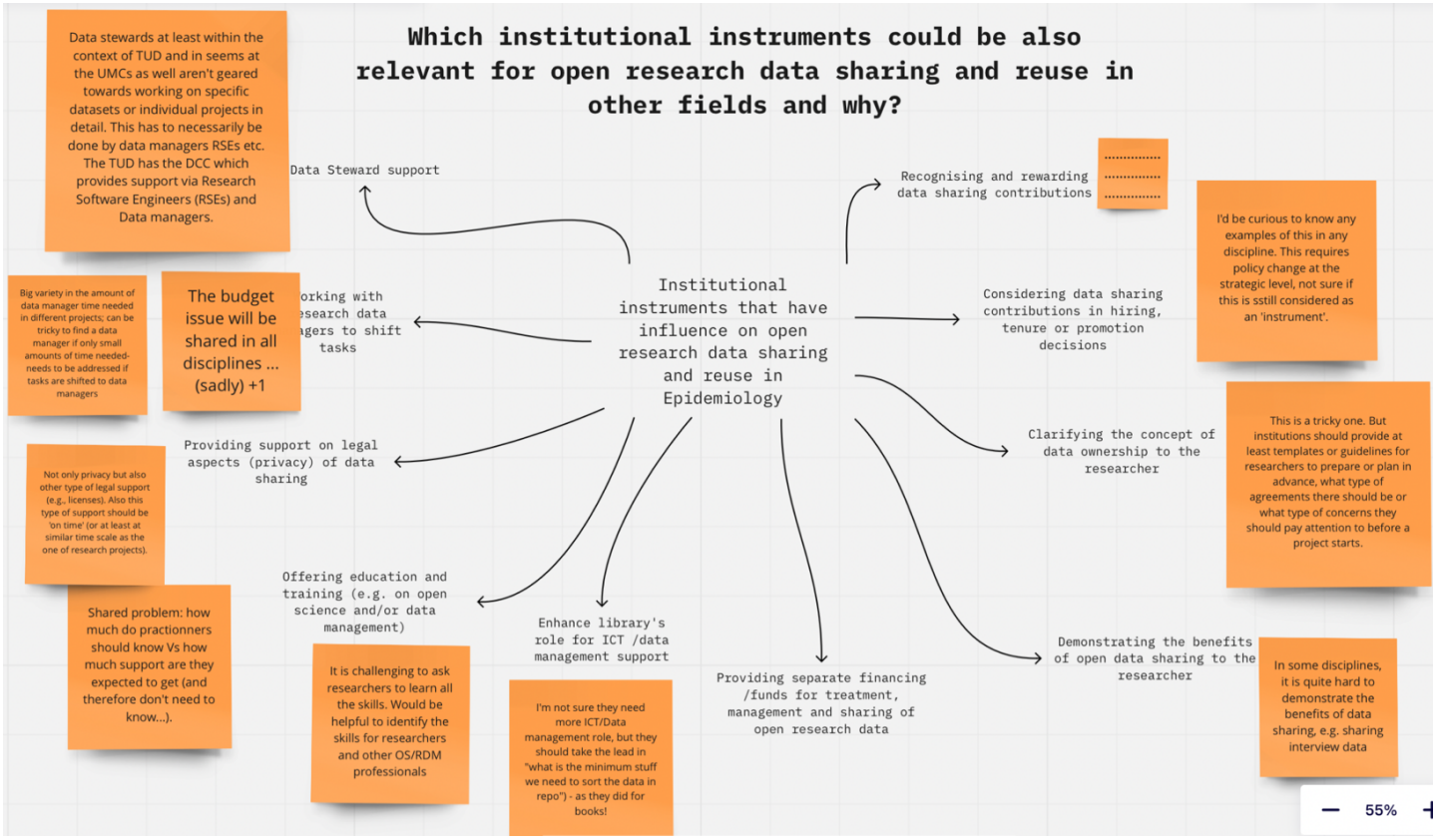

Figure 16 Second activity: gathering insights on institutional instruments

5.4. The findings of the workshop

5.4.1. Usability of the findings in other fields, suggestions, and criticisms

During the workshop, several points were made that explain how certain findings of our study could also apply to other fields or contexts. Several points were made that provide an explanation for our findings and confirm the results of our case study. Several participants made suggestions on how certain instruments could be further enhanced or how they could be complemented with other instruments. There were also contrasting statements to a few of our findings where participants discussed their doubts about certain instruments, citing possible issues with effectiveness and feasibility. Table 13 provides an overview of these findings. To hide the gender of the participants, the third person singular pronoun "they" is used in the text below when referring to the participants.

Table 13 Overview of the workshop findings

Case Study Findings (In the context of Epidemiology)	**Usability of the finding in other research fields or contexts**	**Suggestions in the context of the case study finding**	**Doubts and criticism towards the case study finding**
The infrastructure's capability to handle a large volume of data (in a timely manner) is important for open data practices. However, Epidemiology researchers struggle with accessing this capability.	-In the field of Geophysics, some instruments generate a large amount of data, which in practice causes problems with findings repositories that accommodate these data cheaply and effectively.		
The lack of good search engines inhibits the findability of research data in Epidemiology.	-This is a problem for the overall scientific community. Currently, all fields suffer from the lack of satisfactory engines.	-Alternative methods should be developed to compensate for the suboptimal search engines.	-Building sufficient search engines may not be possible.
Epidemiology researchers get demotivated from open data practices because the available data repositories are not always easy to use during the process of downloading or uploading data.	-This could be relevant for all the fields that do not necessarily collect primary data but also frequently get data from other sources, because the high number of data resources causes problems with hardships of usage.	-Data infrastructures should converge: "Standardize way of working among different repositories" (similar to the discussion in chapter 4.7.1.4.)	
The open data repository's offering of standardized metadata is an important motivator for open data practices in Epidemiology.	-This is relevant for all the research fields, as this instrument makes the data search easier and it makes the data interoperable with other datasets in all research fields.	-Data archives should allow for disciplinary metadata to be added to the citation metadata, and should provide guidance on how to archive extra metadata items.	
Providing timely, structured support for legal aspects (e.g. privacy) of open data practices is important for open data practices in Epidemiology	-Providing legal help is relevant for all fields that deal with personal data (e.g. such as management). In these fields, legal procedures in practice take a lot of time, and a lack of timely support is indeed a demotivator.	-Legal should help not only focus on only privacy but also on other types of legal issues such as licensing. -Focus on tackling misunderstandings or lack of communication (e.g., legal experts speak a different 'language'	

		compared to that of researchers, or there is a lack of templates).	
Offering educational support and training on technical aspects of data sharing is an important factor for open data practices in Epidemiology	- In practice trainings on (open) data sharing prove to be useful in other fields such as Electrical Engineering, Mathematics, and Computer Science. -In practice, trainings prove useful when the background of the researcher does not match the requirements of the field in the context of data sharing.	-Focus on clearly identifying the skills that researchers should have and the skills that other RDM professionals should have -Establish a "common base" of education with standardized curricula	The effectiveness of educational support should not be overestimated: it is hard to provide organizational-level educational support because every researcher or every research project has different situations, and the need to be fulfilled by trainings is very different.
Demonstrating the benefits of data sharing is important for open data practices in Epidemiology.			-Demonstrating the benefits of data sharing would be hard to achieve especially in the case of qualitative research and qualitative data.
Epidemiology researchers expect data stewards to work on specific datasets or individual projects in detail, although traditionally these roles are not taken by data stewards.	- In practice, the role of a data steward is not geared towards working on specific datasets or individual projects in detail, and these responsibilities are to be taken by data managers in research organizations.	-Build a clear distinction between roles (e.g. between data managers and data stewards), so that researchers know who to refer to when they need help.	
There is not enough budget for enabling Epidemiology researchers to work with data managers in the context of open data sharing.	-The (lack of) sufficient budget issue is an issue in other disciplines: only in the fields where data sharing is really common (e.g. genomics research), there is usually a budget allocated for data sharing purposes.	- Create a working system that allows researchers to work with a data manager as much as they need, even if only small amounts of time are needed.	

Handling large volumes of data could be a challenge also for other fields

Regarding the data infrastructures' capability to handle a large volume of data (i.e., large enough for the field), it was discussed that in the field of geophysics, certain types of instruments also generate large volumes of data (e.g. 10 to 40 TB of data for one study), and although there are repositories that could handle datasets of that size, these repositories are extremely costly. This also suggests that the barrier related to handling big chunks of data for data sharing practices also has a relation to the financial struggles of data sharing practices.

The lack of powerful search engines on data repositories could be a general problem

The discussion in the workshop confirms our findings on how the lack of good search engines inhibits the findability of research data. Moreover, the discussion suggests that the issue of not having powerful search engines on data platforms could be a general problem among fields. A participant pointed out that powerful search engines do not exist on repositories, regardless of the field. More importantly, the participant stated that they do not think that such search engines will come any time soon as they believe these engines are very hard to build. It was mentioned by the same participant that providers should then look for other strategies that would compensate for these suboptimal search engines on the platforms.

Ease of use of data infrastructures could be achieved by converging infrastructures

A participant discussed that for researchers such as Epidemiologist -who do not necessarily collect primary data but also frequently get data from other sources such as hospitals and GPs- the ease of data infrastructures could be an important factor for data practices because relying on different data sources means that the researcher needs to be able to handle different infrastructures that correspond to these resources. This participant suggested that to address this issue, data infrastructures should converge, which means that they should be similar or the same to each other in terms of user experience. This is in line with one of the infrastructural instruments that emerged from our case study (which we labeled as "Standardize way of working among different repositories" in chapter 4.7.1.4.): if there is a standard way of engaging with a data repository or a source, then this could have positive influences on open data practices.

Participants agreed on the importance of offering standardized metadata on data platforms

The workshop discussions confirmed our findings on the value of offering standardized metadata on data repositories and archives. It was discussed that data archives should allow for disciplinary metadata to be added to the citation metadata, or should at least facilitate the process (by providing guidance) of archiving extra metadata items in, for example, XML files. The participants agreed that it is important for archives to have disciplinary metadata standards that can be added as extra metadata items in the repository because it makes the searches easier and it makes the data interoperable with other datasets. Furthermore, it was also suggested that the research communities themselves should develop and agree on disciplinary metadata standards.

Support for legal aspects of data sharing is not limited to the privacy aspect

The participants confirmed the value of offering help to researchers concerning the legal aspects of data sharing. More importantly, our findings on the importance of giving timely legal help (e.g. not being too late to communicate to the researchers the requirements of informed consent) were confirmed by the participants, who stated that legal support needs to be 'on time' (or at least at similar time scale as that of the research projects). A participant shared their experience that legal procedures can indeed take a lot of time, and that delays in open data sharing are frustrating for researchers. The participant added that in practice, delays happen because of misunderstandings or lack of communication (e.g., legal experts speak a different 'language' compared to that of researchers, or there is not even a template that could make the process more efficient). This discussion supports our findings that imply that the communication between the legal teams and researchers should be restructured to improve open data practices.

Furthermore, a participant pointed out that legal help should not only focus on privacy but also on other types of (legal) support such as licensing, although our case study analysis did not discover that researchers struggle with licensing issues. It was mentioned during the workshop discussion that support for open content licenses could be important, although these licenses

do not necessarily apply to all the types of data that researchers want to share. The participant pointed out that if there was more support for drafting special licenses for certain types of data it would be much clearer for the data reuser under what terms they can reuse these datasets.

Deciding on how much knowledge and skills researchers should be given is a problem

Offering educational support and training on technical aspects of data sharing was a central discussion topic in the workshop. Although the participants supported the importance of providing institutional support in training researchers to engage in data practices, many of them brought attention to the problem of organizations not being able to understand how much technical knowledge they can ask researchers to learn. The issue is about the inability to decide on the level of skills and knowledge the researchers should be given by trainings. A participant pointed out that the shared problem across fields is that currently, the line is not clearly drawn between *how much researchers should learn themselves (via trainings)* versus *how much complementary support they should be given (by supporting agents like data managers and data stewards) to cover the rest.* Understanding to what extent the educational support aims to train researchers is valuable because, without this being settled, the complementary support (from data managers, and data stewards) cannot efficiently be formulated and adjusted to help researchers. The participants noted that this "split" is not talked about much in their institution: the institution gives researchers training on everything and expects researchers to know everything (i.e., towards building the maximum level of capabilities). However, there is also support staff that aims to assist with everything, which signals that the institution could be too harsh on their expectations from researchers regarding the level of technical capabilities they should have. The participants, therefore, suggested clearly identifying the skills that researchers should have and the skills that other research data management (RDM) professionals should have. Based on this, organizations should put effort into giving researchers a "common base" with standardized curricula, and then aim to give advanced trainings, if needed, based on the individual needs of researchers.

The participants discussed that it is hard to provide organizational-level educational support (on technical aspects of data practices) because every researcher or every research project has different situations, and the needs to be met through training are very different. It was stated that in practice, any kind of educational support at the institutional level could cover a maximum of around 80%, but then the other 20% should be customized help. The participants believe it might be useful to adjust the expectations, and steer away from expecting institutional educational support to be able to solve all the problems related to researcher capabilities and skills in data practices. There needs to be a combination of different elements (support from data managers, data stewards, trainings, etc.), and even then, there is still effort needed to solve individual cases which ask for unique technical capabilities on data practices. The attention given to combining different institutional instruments confirms our research findings that there is benefit in combining instruments to ensure effectiveness and the best outcomes.

Demonstrating the benefits of data sharing could be tough, especially for qualitative research

The participants somewhat agreed on the value of explaining the benefits of data sharing to the researchers, however, they noted that in practice this would be hard to achieve especially in the case of qualitative research. A participant shared an observation that in qualitative research (e.g. that includes interviews or focus groups as the data collection), the data sharer cannot really see the benefits in data sharing, because the data are not cited or reused by others immediately after the sharing of data. Therefore, the participants doubted the feasibility of this instrument in practice.

It is important to highlight the distinction between the roles of data stewards and data managers

Regarding our discussion on Epidemiology researchers' struggle of reaching data stewards for individual help (such as on data anonymization), the participants discussed that in their organization, the role of a data steward is not geared towards working on specific datasets or individual projects in detail, and that these responsibilities are to be taken by data managers. The participants discussed the importance of building a clear distinction between roles so that researchers know who to refer to when they need help. Our case study analysis indicates that, in the universities that we examined, this role division may not be clear, which causes Epidemiology researchers to build unfeasible expectations from data stewards in their organization.

Not having a budget for working with data managers is a general problem across fields

During the workshop, we obtained insights on how the lack of financial resources for data sharing practices is not just an issue for Epidemiology. The participants mentioned that the lack of sufficient budget is possibly an issue in many other disciplines, citing that only in the fields where data sharing is really common there is usually a budget allocated for data sharing purposes. Regarding being able to work with data managers to shift tasks, a participant brought attention to the fact that there is a big variation in the amount of data manager time needed in different projects. Some researchers may need a data manager for a long period, while some may require a very little amount of time from them. The participants noted that finding a data manager if only a small amount of time is needed could be tricky. So, it is an organizational responsibility to create a working system that allows researchers to work with a data manager as much as they need, even if only small amounts of time are needed.

5.4.2. Prioritization of institutional instruments over infrastructural instruments

The survey data we obtained give us insights into which instruments are perceived to be more important by the participants. The results show that institutional instruments can play a stronger role in enhancing open research data practices, which confirms our previous arguments in chapter 4.7.4. In Figure 18, it can be seen that the participants of the workshop favored institutional instruments over infrastructural instruments.

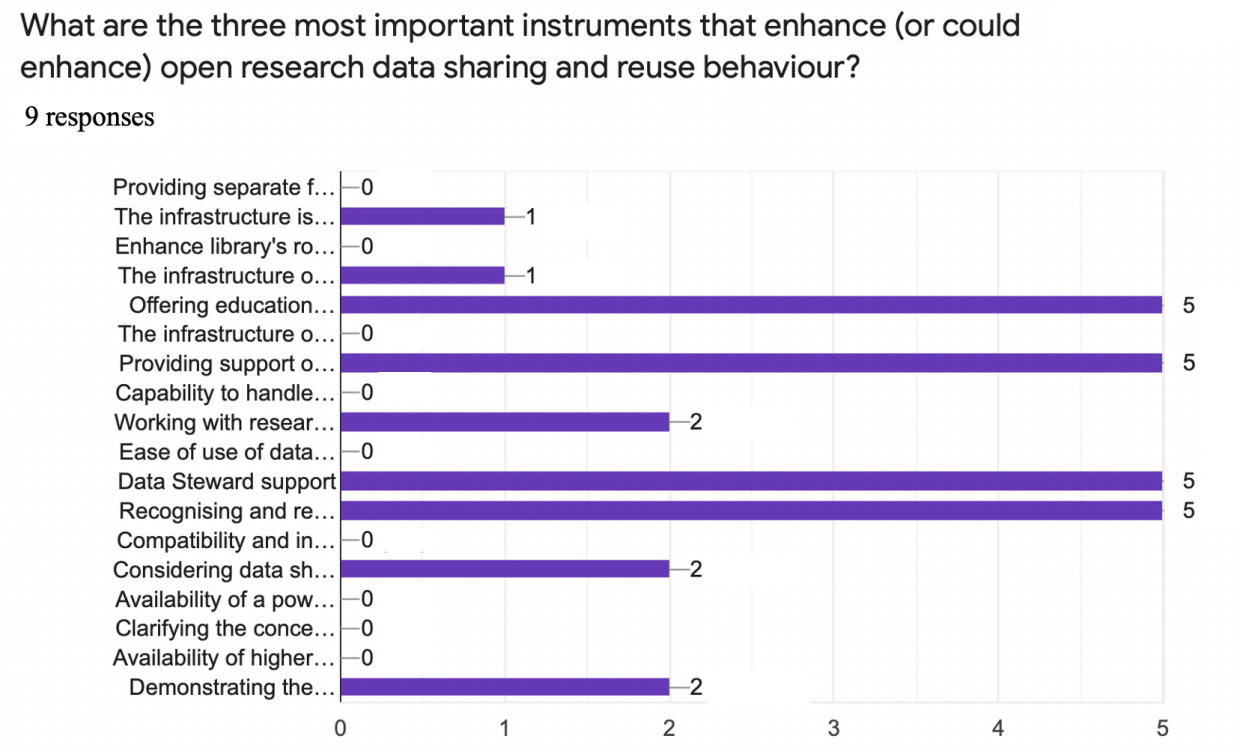

Figure 18 Responses to the survey

The instruments that were selected the most are *offering education and training (e.g. on open science and/or data management)*; *providing support on legal aspects (privacy) of data sharing*; *data steward support*; and *recognizing and rewarding data sharing contributions.* These four instruments were all selected five times (see Figure 18). Several participants also explained the reasoning behind their choices. A participant explained that the instruments that are more "direct" (i.e. the ones that provide immediate support) should be prioritized, which - in that participant's view- would be providing support on legal aspects and data steward support. Similarly, another participant also stated data steward support (and also working with data managers) is important because such agents provide "hands-on" support for data practices. The participants stated that data steward support can accomplish the realization of many other instruments (that we mentioned) since data stewards are facilitators and policy advisors who play the role of "operationalizing" things.

Another participant explained that training is important because it helps to have someone show researchers how to openly share data -despite open data sharing is, in many cases, not hard. Another participant explained that legal issues are possibly one of the hardest to deal with, which is why professional legal help is valuable. The instrument of recognizing rewards was chosen by some participants because these participants found this instrument to be a strong motivator for researchers (e.g. it will also incentivize researchers to learn how to share data). All these explanations confirm the findings of our case study.

Finally, it shall be noted that this workshop's participants, who are mostly data stewards, all work on the institutional side of the problem that we evaluated in this case study. Therefore, the fact that they ranked the institutional instruments over infrastructural ones could also be related to the background of the participants. On the other hand, because these professionals collaborate closely with colleagues working on the infrastructural side, they still provided many insights into the role of infrastructural instruments and their potential across fields and contexts.

6. Recommendations

Based on the case study analysis and the workshop that were described in the previous chapters, this chapter provides recommendations to infrastructure developers/providers and management/policymakers of universities so that open data sharing and reuse practices in Epidemiology can be enhanced. We derived these recommendations by evaluating the instruments that were found to be important but lacking (e.g. in terms of existence or the current use structure) and by examining the suggestions given by the participants (i.e. the researchers, the research data management consultant, the data stewards, and the data officers) in our study. These recommendations include many changes such as enhancing functionalities in data repositories and revisioning certain roles and structures in organizations. Table 14 provides an overview of these recommendations.

Table 14 Overview of the recommendations

Actor category	**Recommendation Category**	**Recommendation**
Infrastructure developers	Enhance findability of research data by expanding features	Infrastructure developers should build search engines on open data repositories allowing advanced search options (similar to the search engines that exist in repositories where researchers look for publications, such as PubMed)
		Infrastructure developers should build technical functionality to differentiate scientifically relevant versus "invaluable" or "junk" datasets.
		Infrastructure developers should look for alternative strategies that would compensate for suboptimal search engines on the platforms.
	Consider metadata and data dictionaries as obligatory elements of data sharing	Infrastructure providers should steer from the current soft policies that state "add as much documentation and metadata along with the datasets as possible" towards harder policies.
		The research community should develop and promote standardized disciplinary data dictionaries.
		Infrastructure providers should make sure that data repositories explicitly ask users to adhere to discipline-specific guidelines (not general guidelines) during data upload processes.
	Invest in infrastructures that tackle privacy concerns	Infrastructure developers should build novel data infrastructures that, by design, could provide a workaround to the privacy concerns (i.e. by enabling users to run code without accessing data, or by enabling users to access synthetic data).
		For the Netherlands to benefit from the conveniences brought by technical developments; infrastructure developers, the government, and universities should develop roadmaps for building systems similar to OpenSafely for the Netherlands.
University (management) and policymakers	Restructure the communication between legal teams and researchers	University policymakers and management should streamline the communication process so that researchers get valuable legal knowledge on time.
	Clarify the role of data stewards and consider their central figure	University policymakers and management should position and promote data stewards to researchers as "the" point of contact when they need advice and guidance on services.

		University policymakers and management should divide the roles of data stewards and the roles of data managers clearly (and consider more funding for data managers).
		University policymakers and management should build concrete, separate roles and responsibilities for data governance in the organization, and build a roadmap that guides researchers on what they can do in case they have not been able to work with data managers but still need support, for example, on opening up datasets.
	Establish what is expected from researchers and what is expected from other researchers (in terms of capabilities)	University management should identify the skills that researchers are expected to have and the skills that other RDM professionals should have so that these professionals can support researchers efficiently.
	Consider the library's potential for ICT support for the future	University management should structure the positioning of the library considering the future demand for ICT services on data sharing.
	Focus on field-specific open science curricula targeted for all researchers	University policymakers and departments should prioritize institutional instruments, especially the ones that contribute to building a culture of data sharing.
		University management should develop communication channels (other than policy documents) to explain the benefits of open research data sharing.
		University management and departments should ensure that open science trainings are also given to more senior researchers (not just the PhDs).
		Open science communities and university management should build a stronger "field-specific" educational curricula to give researchers a hands-on skillset that they can apply to their research activities.
		Departments should ensure that Epidemiology researchers are taught how to make data dictionaries that are in line with the needs of the Epidemiology field.
		Departments should ensure that Epidemiology researchers are trained on how to use Github for sharing research codes.
	Re-frame the data ownership concept and focus on data controllership	University policymakers and management should re-evaluate the framing of the data ownership concept towards establishing both themselves and researchers as agents who are controlling data.
		Open science communities and universities should make sure that conversations about data ownership are included in the open science curricula.
	Make open research data contributions as a prominent part of Reward and Recognition Programs	Universities and departments should add open data sharing contributions as a key pillar (not just a sub-criterion) to their general principles for the assessment of research in the Rewards and Recognition systems.
	Separate the policy for open metadata sharing from the policy for open research data sharing	University policymakers should consider putting a stronger emphasis on policies asking to make metadata (both project-level metadata and dataset-level metadata) open, even when the associated research data cannot be made open.

6.1. Recommendations for infrastructure developers and providers

Enhance findability of research data by expanding features

As the amount of research data is continuously growing, infrastructure developers should ensure that open data infrastructures are adapted to the requirements of big data. For an infrastructure, this does not only mean being able to accommodate large scales of data (which over the last decade has been made possible by cloud-based technologies that currently support many repositories), but also dealing with the issue of findability of the data in the midst of a "data" boom that is defining the 21st century. Our case study suggests that Epidemiology researchers get demotivated by their engagement with open data infrastructures, particularly during the process of searching for data. There are several points of concern. The first is not being able to perform complex search queries**. Our first recommendation is that search engines on open data repositories should have advanced search options similar to the search engines that exist for repositories where researchers look for publications, such as PubMed**. The second point of concern relates to constantly getting lost in irrelevant, scientifically invaluable datasets when a data search is performed on a data platform. In theory, this problem should be solved by ensuring that data uploaders adhere to standards ensuring interoperability (i.e. by adhering to using metadata standards, building standard data dictionaries, etc). However, our case study suggests that in practice, this is not the reality, and that there is a vast number of datasets on data repositories whose value for reuse ranges from having none to being highly valuable. For an infrastructure system designer, these issues have implications on the system requirements: **Search systems or data repositories should be able to differentiate scientifically relevant versus "invaluable" or "junk" datasets, because this responsibility is too difficult to be taken by the user in the era of Big Data, where a simple search with a keyword brings thousands of results. We recommend developers to look for methods to give platforms functionality for differentiating data according to (re)use value**. As building powerful search engines could be a tough task, developers should look for alternative strategies that can address suboptimal search engines.

Consider metadata and data dictionaries as obligatory elements of data sharing

It is also worth mentioning that for Epidemiology or public health fields in general, the value of data depends significantly on how comprehensive the data description is. We argue that the decision to put a comprehensive data dictionary along with the dataset should not solely lie with the user (i.e. the researcher who is uploading the datasets), but the infrastructure design should also put forward a hard constraint that inflicts this. **The infrastructure policies should therefore reconsider the current soft policies implying *"add as much documentation and metadata along with the datasets as possible"***, which, for instance, currently forms the guideline of Dataverse.nl: "*At a minimum you need a title for the Dataset, the name and affiliation of the author(s) or data collectors, subject and a description about the data. If you have more information about your data, you can fill in other metadata fields. Also, you can upload documentation files and code files to accompany your data files. (This is not obligatory.) The more relevant information you provide, the better your dataset will be*

findable." (*DataverseNL: FAQ*, n.d.). Such a policy may be just sufficient for other research fields, but for fields that are highly sensitive to metadata on data collection and data characteristics, such as the field of Epidemiology, we recommend upholding more extensive requests. We argue that in Epidemiology, having extensive metadata and a detailed data dictionary explaining how the data have been collected is not something that should be evaluated as complementary material, but as a "must" for data sharing. Currently, on data repositories, the amount of information on data collection and the way such information is presented varies heavily from one dataset to another. **Our analysis shows that the guidelines on making data dictionaries should then be standardized for the fields of public health**, and these guidelines should be promoted both by the infrastructures and organizations. The standard guidelines from data infrastructures (such as that of OSF (*How to Make a Data Dictionary*, n.d.)) for preparing data dictionaries do not suffice for Epidemiology because they are not detailed enough. **We recommend data infrastructures to explicitly ask users to adhere to field-specific guidelines (see Bialke et al. (2015) for a notable example) during data upload processes.**

Invest in infrastructures that enable running code on data without accessing the data

We suggest that there should be a focus on developing and adopting novel data infrastructures that, by design, could provide a workaround to privacy concerns. Since open metadata sharing is always a possibility (since it is not bounded by the GDPR), infrastructures that allow for deidentified and secure processing of data -without sharing the data themselves- have the potential of being highly valuable for the fields of public health if they are adopted on a higher level (Mitchell et al., 2021). A notable initiative is the OpenSAFELY platform in the UK that allows secure analysis of public health records (Mitchell et al., 2021). Even though it is questionable whether infrastructures like this would fall under the definition of "open" data infrastructures, we believe they can still result in major changes in the level of data sharing in the field of Epidemiology where data sharing occurs currently mostly by collaboration and one-on-one data exchange. **We recommend developers in the Netherlands to develop roadmaps for similar initiatives, and also focus technically on which other types of research data these data infrastructures can accommodate other than public health records.**

6.2. Recommendations for university managements and policymakers

Restructure the communication between legal teams and researchers

Researchers currently view legal teams and data privacy officers as agents who are somewhat "against" open data sharing practices. We observe that if researchers go to privacy officers at the end of the research process to make their research data open, there is a higher chance for the application of (open) data sharing to be rejected due to privacy concerns. In practice, applications of open data sharing get rejected simply because researchers were not aware of the exact informed consent procedure that they should have adhered to at the beginning of the research cycle. Data management plans may not be sufficient tools to communicate this, as researchers find them too long and burdening. **We recommend universities to focus on**

streamlining the communication between people who are responsible for handling privacy issues (i.e. data privacy officers) and researchers. Our analysis suggests that engagement between these parties could turn to more positive outcomes if the information exchanges happen earlier in the research cycle. If data privacy officers can communicate early in the process what exactly needs to be done to comply with the privacy regulations, there are more possibilities for open data sharing compared to the case where a researcher consults a data officer after the research ends -which is long after the data collection.

Clarify the role of data stewards and consider their central figure

In the Netherlands, data stewardship approaches differ by university. For example, UMC Utrecht's approach is more on a departmental level, aiming to give "tailored support" for project-specific contexts (Böhmer, 2019), whereas other universities may opt for faculty-level approaches. Regardless, data stewardship (role) is valuable for open data practices especially because it is the contact point between the researchers and many services such as those related to ICT, libraries, HR, finance, and legal aspects of research data governance. Our analysis indicates that data steward services may be suffering from a lack of visibility. **We recommend university departments to position and promote data stewards to researchers as *"the"* point of contact when they need advice and guidance on services.** Furthermore, our analysis shows that researchers feel as if data stewards are too busy to help them in more complex cases. We argue that, apart from the issue relating to being understaffed, this issue could also be due to the roles of data stewards and data managers not being properly differentiated in the UMCs. When researchers are not able to work with a data manager, researchers may expect to have these services from data stewards. This suggests a structural problem in the universities stemming from the issues of (1) data managers being hard to reach due to financial unavailability and (2) the role of data stewards not being clearly defined. For the first issue, **we recommend universities to provide more financial funds for researchers to be able to work with data managers. We suggest looking for alternative funding strategies for data managers since funds from grants are not sufficient in practice**. We observed that in some universities, data managers are hired on a departmental level. This could be a promising strategy that other organizations can adopt so that researchers who have relatively smaller projects could still have access to working with data managers. For the latter, **we recommend universities to formulate concrete roles and responsibilities for data governance, define the concrete role of data stewards to researchers, and build a roadmap that guides researchers on what they should do in case they have not been able to work with data managers but still need support, for example, on opening up datasets.**

Consider the library's potential for ICT support for the future

Many researchers have expressed having no engagements with libraries on topics of research data management before. We recommend universities to position the role of libraries in the context of research data management support more concretely. **We argue that libraries could be a supporting agent, especially for ICT support since researchers expressed frustrations over the lack of technical help in their past experiences with data sharing.** Universities and policymakers should pay attention to making long-term investments regarding how they will

provide technical support to researchers in the future as it is likely that there will be more open data practices due to open science adoption. If providing more funds for working data managers or research data management is a utopia, universities have to make strategic decisions that lead other agents, such as university libraries, to prepare for the future demand for these services.

Focus on field-specific open science curricula targeted for all researchers

Our analysis indicates that universities should prioritize instruments that would contribute to building a culture of data sharing. Policy documents are not read or superficially read by researchers. This means that researchers need to be communicated via other channels to gain the necessary knowledge and perspective that the policy documents are indenting to give towards open data practices and the open science ideology (e.g. making data FAIR). Since understanding the benefits of open data practices is considered to strongly influence open data sharing and reuse, universities need to be able to communicate these to researchers in novel approaches other than solely relying on the policy**. We recommend university boards and department heads to look for methods of communicating these ideas. In that regard, trainings in open science are useful for researchers to understand the benefits of open research data sharing**. Our analysis supports the current practices of providing concrete open science curricula in universities because, in the long run, it ensures everybody gets sufficient education. However, currently, university open science roadmaps (such as that of Utrecht University), pay attention to giving open science trainings only to master and PhD students (de Knecht et al., 2021). **We recommend the universities to focus on giving open science trainings also to more senior researchers**, because senior researchers have the executive power to openly share research data, and they are relatively far from open science culture (in terms of mindset). We recommend these modules to be required for all researchers. Second, **we advise strengthening the "field-specific" open science curricula to better communicate the benefits of open data practices in the context of the field, and also to give researchers a hands-on skillset that they can apply to their research**. For example, making a detailed data dictionary is considered to have great value in open data sharing in Epidemiology, whereas this may be less of an issue for other fields. Thus, the tailor-made open science curricula of Epidemiology should, for instance, provide more emphasis on this aspect. **We recommend universities to teach Epidemiology researchers how to make data dictionaries that are in line with the needs of the Epidemiology field as part of a required (not voluntary) curriculum**.

Re-frame the data ownership concept and focus on data controllership

We advise university boards and policymakers to re-evaluate the framing of the data ownership concept. Currently, researchers may intrinsically believe that they are the sole owners of data as they have gone through the efforts of collecting it. When someone feels like they "own" something, the act of "giving it away" or "letting it go" could have negative connotations. Although the owners of the data are mostly universities by law, we advise that the conversations regarding data ownership should not be steered towards conversations of *"it is not your data, but mine",* but rather conversations where scientific knowledge (e.g. research data) is framed as a public good. **Universities should frame both themselves and researchers**

as agents who are controlling data and who have the responsibility to make data accessible to the public as much as possible. These conversations about data ownership should be integrated into the open science curricula.

Make open research data contributions as a prominent part of Reward and Recognition Programs

One of the leading barriers in front of open data sharing is the lack of a reward or recognition system for open research data contributions. Our study indicates the potential for the *Rewards and Recognition* programs that several Dutch universities have begun implementing in recent years. As guided by the San Francisco Declaration on Research Assessment (DORA), these programs intend to steer away from conventional metrics such as journal impact metrics when evaluating a researcher's performance, and instead value scientific products by creating a mix of qualitative and quantitative indicators as well as narratives (*Recognition & Rewards - Research*, n.d.; *TU Delf Recognition & Rewards Perspective (2021-2024)*, n.d.). For example, Maastricht University recommends looking for a vast number of different criteria, such as output indicators, amount of external funding, number of PhD students supervised, recognition of research (prices won), scientific integrity, degree of influence in the organization and research community, etc. (*Recognition & Rewards - Research*, n.d.). Our analysis indicates that if open research data contributions are considered in hiring, promotion, or tenure decisions, researchers would be more willing to openly share and reuse research data. Therefore, to ensure that these programs have a positive influence on open data practices, **we recommend universities or departments to add open research data sharing contributions as a key pillar (not just a sub-criterion) to their general principles for research assessment in the Rewards and Recognition programs**.

Separate the policy for open metadata sharing from the policy for open research data sharing

An instrument that was brought up during our study is increasing the communication within the broader scientific community to help researchers be aware of other researchers who are conducting similar research (and therefore producing similar research data). We believe the communication among the scientific community could be influenced heavily by openly sharing metadata. Our case study analysis found no reason why Epidemiological research metadata should not (or cannot) be publicly available at all times. Availability of metadata in the public domain would allow researchers to be aware of the existence of datasets even at times the primary data cannot be made public. We argue that there could be a policy issue that needs to be addressed by universities: organizations and policy documents promote the act of open research data sharing as sharing of all kinds of research data. They do not make a clear distinction for sharing metadata, and they do not mention the value of metadata sharing at times sharing the primary research data is not a possibility. **We recommend organizations and policymakers to consider putting a stronger emphasis on their documents and operations to incentivize making metadata (both project-level metadata and dataset-level metadata) open,** apart from encouraging making research data "as open as possible and as closed as necessary" as part of the current FAIR data guidelines (Landi et al., 2020).

7. Conclusion

This chapter first summarizes the motivation of the research and provides answers to the main research question as well as the associated subquestions of the study. Moreover, it discusses the research project's scientific contributions, societal relevance, and suitability to the Complex Systems Engineering and Management (CoSEM) MSc program. Finally, it discusses the limitations of the study and gives directions for future research.

7.1. Motivations and main research question

Advancements in communication technologies over the last decades have led to an increase in the amount of data that are collected, analyzed, and stored by scientific communities (Tenopir et al., 2011). As science is shifting toward data-driven research, research data sharing also gains utmost importance (Hey et al., 2009). The data that researchers collect, process, and analyze during the research cycle can continue to create value beyond the primary research publication through research data sharing practices. Especially sharing research data "openly", which refers to publishing on the internet in a freely accessible, usable, modifiable, and sharable format to other researchers, has significant benefits to researchers and scientific fields (Zuiderwijk & Spiers, 2019, p. 229). These benefits range from increased transparency in the research (Patel, 2016) to decreased researcher time and effort on repetitive and unnecessary data collection processes (Tenopir et al., 2011). Despite the benefits, there are factors limiting open research data sharing and reuse practices such as lack of data standardization and lack of time (van Roode et al., 2018). Previous research has done in-depth examinations on both the drivers and inhibitors of data sharing motivations (Tenopir et al., 2011; Zuiderwijk et al., 2020; Zuiderwijk & Spiers, 2019). Acknowledging that the issues limiting open research data adoption differ heavily by field, it is important to do field-specific studies to understand the discipline-specific, contextual challenges as well as opportunities for promoting open research data practices. Since the field of Epidemiology is considered to have lower levels of data sharing practices, there is a possibility for significant value creation if ways for promoting open research data sharing and reuse are established in this field. Although many factors contribute to researchers' open research data sharing and reuse motivations, focusing specifically on infrastructural and institutional issues and possibilities could give infrastructure developers, policymakers, and research organizations better guidance on how these practices can be enhanced.

In line with the motivations above, the main research question of this study is "*What roles can infrastructural and institutional arrangements play in promoting open research data sharing and use behavior in Epidemiology?"*. To answer this question, four subquestions were formulated and subsequently answered using distinct qualitative research methods.

7.2. Answering the subquestions of the study

Subquestion 1: What infrastructural and institutional instruments influence researchers to openly share their research data and to use openly available research data?

This subquestion investigates the infrastructural and institutional instruments that could influence open research data adoption via performing a systematic review. We formulated the first subquestion not specific to a domain but across all research domains for two reasons: (1) we acknowledged that there is no specific line of literature on instruments enhancing open research data adoption in the field of Epidemiology, and (2) we proposed that it is valuable to examine those instruments that prove useful in other domains in our Epidemiology-focused case study. The procedure of the systematic literature review was established in chapter 3.2. This literature review resulted in a range of infrastructural and institutional instruments.

We classified the infrastructural instruments into three categories: (1) *instruments enhancing the usability of the infrastructures,* (2) *instruments supporting the facilitation of FAIR data principles,* and (3) *instruments concerning security and trust aspects*. Regarding usability of infrastructures, we illustrated that open data infrastructures such as open data repositories should have easy-to-use interfaces, should be able to handle a large volume of data, should be reliable, should be integrated and compatible with other infrastructures (such as analysis software), should allow for data analysis (as an integrated feature), and should offer assistance for the choice of repositories as well as licensing issues. Regarding instruments relating to FAIR principles, we illustrated the importance of metadata standards, infrastructures' capability of storing and showing metadata, showing data usage statistics on repositories, providing powerful search engines, and offering assistance for data citation. Furthermore, we illustrated that individual data repositories should be linked with overarching registries and that infrastructures should be compatible with domain-specific requirements relating to data. Regarding the third category, we noted data anonymization tools as an instrument.

Based on our findings, we classified the institutional instruments into four categories: (1) *instruments governing the data sharing and reuse process*, (2) *instruments actively supporting researchers in the data sharing and reuse process,* (3) *instruments providing financial resources,* and (4) *instruments that build a culture of data sharing and create incentives.* Regarding the first category, we noted providing institutional data management and data sharing policies, as well as enhancing the practice of data management plans as instruments. For the second category, we found out that guiding the researchers to select an appropriate repository, providing support from data stewards and legal teams, and providing possibilities of working with research data managers could be some of the instruments. For the third category, we noted the instrument of providing separate funds for research data management. Some of the instruments in the last category are recognizing and rewarding data sharing contributions, demonstrating the benefits of data sharing, and requests for open data sharing.

The findings of the literature review resulted in the formation of the conceptual framework (chapter 3.3.), which is provided as input to the case study in chapter 4.

Subquestion 2: How do infrastructural and institutional instruments influence researchers in openly sharing their research data and in using openly available research data in the field of Epidemiology?

This subquestion refers to the examination of the instruments that influence Epidemiology researchers' motivations for open research data sharing and reuse. We chose the case study research approach since our research question is interested in "how" open research data practices occur the way they occur, and also because the research nature is highly exploratory due to complex field-specific dynamics. The main information source of the case study was ten Epidemiology researchers who work in various research institutions in the Netherlands and a research data management consultant working in one of these institutions. We systematically coded the interviews to analyze them, documented our analysis in the codebook, and illustrated our operationalization process. In the interviews, we focused on understanding (1) whether the instruments we examined were available to the researchers (availability), and (2) the extent to which these instruments would influence open research data sharing and reuse (importance).

Before analyzing the instruments, we described the important characteristics of Epidemiology related to our research topic. For example, we found out that Epidemiology researchers deal with large datasets which they often collect from cohort studies that could take years to conduct, which makes data collection hard. We noted how clinical work puts extra time pressure on Epidemiologists compared to other fields. We explained that obtaining informed consent from patients in a clinical context is relatively harder. We illustrated that research agendas in Epidemiology could be very flexible: researchers could develop research questions at various stages of the data collection process. We also demonstrated how the GDPR could be inhibiting the data sharing practices in the fields of public health, and how data anonymization could be much less powerful -contrary to our literature findings. We described the prevalence of certain data sharing types such as data sharing by collaboration and one-on-one data sharing, while also bringing attention to the lack of an open data sharing culture in this field.

Several infrastructural instruments are perceived to be highly important for open data practices in our case: (1) easy-to-use interfaces, (2) compatibility between different data infrastructures, (3) availability of powerful search engines, (4) availability of overarching registry of repositories, (5) infrastructure's offering of metadata on data collection, and (6) the infrastructure's compatibility with the domain-specific privacy requirements. For example, Epidemiology researchers could likely struggle with the ease of use of data repositories, especially during the process of uploading datasets. Based on the previous experiences of researchers, we indicated that the complexity of engaging with the data repository could be a demotivator. Moreover, our analysis showed that when working with relatively more complicated data infrastructures such as GitHub or when dealing with less compatible data types, researchers may need technical skills that they most likely did not have the opportunity to acquire during their studies. We also showed that for Epidemiology researchers, being able to find detailed descriptions of how the data were collected is an important motivator for reusing open research data. Furthermore, regarding the findability of research data on data repositories, Epidemiology researchers want to be able to use advanced search queries, similar

to those on search engines like PubMed. The lack of an overarching registry that connects all the repositories in the field also has a negative influence on motivations for open research data reuse.

Several institutional instruments are perceived to be highly important for open data practices in our case: (1) Data steward support, (2) working with research data managers, (3) providing support for legal aspects relating to open data practices, and (4) recognizing and rewarding open research data sharing contributions. For instance, the role of data stewards is valuable because they are the point of contact when researchers have questions about research data management. However, Epidemiology researchers report that data stewards often do not have enough time to help them thoroughly in more time-consuming data sharing activities. Furthermore, being able to work with research data managers could have a positive influence on open research data adoption in Epidemiology, because many researchers perceive that they do not have sufficient time to prepare and maintain datasets for open access due to extra clinical work pressure. We found that only researchers with larger projects get the chance to work with data managers, and even when a data manager is working on a project, data managers do not have time for making datasets open in practice. Our analysis also indicates that there could be communication issues between the legal departments and Epidemiology researchers: data officers have the reputation of being too strict towards open data sharing applications due to privacy considerations. Furthermore, researchers would be incentivized toward open data practices if they were to believe that these efforts are sufficiently recognized and rewarded in the field.

Subquestion 3: To what extent can the case study findings on infrastructural and institutional instruments be applied to other research fields and the general scientific community?

This subquestion refers to evaluating the extent to which the findings of the case can be useful in other research fields. In line with this, we held a workshop with nine participants who either work as a data steward or a research data officer at different faculties of a Dutch research university to examine the transferability of our case study results. We found that many findings of our case study could also apply to other research fields. For instance, we understood that the low findability of research data due to the lack of sufficient search engines is a problem that is experienced by many other technical disciplines. Another common problem in the general scientific community is the lack of financial resources to support researchers with research data management and open data sharing activities, for example, by enabling them to work with data managers. Moreover, we understood that providing timely, structured support for legal aspects of open data practices is important for other research fields that also deal with personal data on a high scale (e.g. policy and management), and that in practice, other fields also struggle with the lack of structured and timely communication between researchers and legal teams in the context of open data sharing. We also found out that a data repository's capability to handle large volumes of data in a timely manner would be valuable also in other fields that use research instruments that conventionally generate significant volumes of data, such as the field of Geosciences. We found out that converging data infrastructures to one another in the context

of user experience during data upload/download would be valuable for all other fields that frequently rely on getting data from secondary sources -instead of collecting primary data by themselves. Furthermore, since all research fields require the shared research data to be interoperable and findable, the value placed on providing standardized metadata across data platforms seems to be common. The results of the workshop also confirmed, regardless of the field, that it is important to clarify the roles of data stewards: it should be established institutionally that data stewards are not geared towards working on individual projects in detail and that these roles should be assumed by other agents in the organization.

The workshop also indicated doubts about some of our findings on infrastructural and institutional instruments. We found out that, although trainings on (open) data sharing in practice prove to be useful in other technical fields (such as Electrical Engineering), the effectiveness of institutional-level trainings should not be overestimated considering that each research project has unique needs. Furthermore, it was indicated that, in the context of qualitative data sharing, changing researcher motivations (e.g. by demonstrating the benefits of data sharing) could be much harder since in practice, researchers practicing qualitative research generally have more difficulties in comprehending the (value of) data sharing concept.

Subquestion 4: How can infrastructural and institutional arrangements in the field of Epidemiology be enhanced so that they are more effective in promoting open research data sharing and reuse?

This subquestion refers to reviewing the findings of the case study analysis and the evaluation workshop, and subsequently discussing what we can learn from this study that would help enhance open research data sharing and reuse practices in Epidemiology. To derive recommendations, we first went over the refined conceptual framework (see chapter 4.7.5.) and then specifically reviewed the points in our case study and workshop where the participants mentioned issues about the application of certain instruments, the need for certain instruments, or what can be changed about certain instruments. We divided our recommendations into two categories: (1) recommendations for data infrastructure developers and providers (chapter 6.1.), and (2) recommendations for universities and policymakers (chapter 6.2.). For the first category, we recommended that the findability of research data should be enhanced on data repositories via expanding features of search engines; that standardized metadata and data dictionaries should be considered obligatory elements of the data sharing process; and that developers in the Netherlands should invest in novel data infrastructures that enable running code on research data without needing access. For the latter category, we recommended universities and policymakers to restructure the communication between legal teams and researchers to ensure that researchers can engage with legal teams from the beginning of the research cycle; to clarify the role of data stewards as the point of contact in data governance support; to consider library's future potential for ICT support on data sharing; to focus on providing field-specific open science curriculum (for researchers from all levels) that gives a hands-on skillset; to reframe the data ownership concept to steer conversations away from conflicts of who "deserves" the rights of data; to make sure that open research data contributions are incorporated in the new Rewards and Recognition systems; and to separate

the policy for open metadata sharing from the policy for open research data sharing to ensure that metadata are openly shared even when research data cannot be.

7.3. Answering the main research question

By providing answers to each subquestion of this study, we accumulated the information that is needed to answer the main research question of this study: **What roles can infrastructural and institutional arrangements play in promoting open research data sharing and use behavior in Epidemiology?**

With this research, we established that researchers in Epidemiology do not openly share or reuse research data due to many different reasons relating to the legal, cultural, technical, and organizational issues. Infrastructural and institutional instruments can enhance open research data practices by addressing these issues. Our research showed that the role and effectiveness of instruments depend on the dynamics of the Epidemiology field. We understood that many of the instruments are not fully within the reach of Epidemiology researchers despite having a huge potential for increasing motivations. For example, researchers do not always have sufficient access to research data managers, data stewards, search engines with satisfactory functionality, overarching registries, or reward systems for research data sharing contributions. Institutional instruments in Epidemiology have the potential to support open research data adoption by reversing the lack of an open data sharing culture with the right incentivization approaches, by the provision of financing, and by actively supporting researchers in the process of openly sharing and reusing research data via engagements with data stewards, research data managers, libraries and data privacy officers. Our research showed that institutional instruments are in a more vital position, so bringing these instruments into use and enhancing them can be prioritized. However, infrastructural instruments also have significant potential for supporting open research data adoption via increasing the findability and interoperability of the research data, and also via making researchers' interaction with various data infrastructures easier considering the number of technical skills needed for open data practices. Some instruments may have little or no role in influencing open research data sharing and reuse behaviors in Epidemiology, such as offering institutional data sharing policies and offering data anonymization tools. Nevertheless, considering that many instruments complement one another, to increase the effectiveness of the instruments in practice, they should be combined. To strengthen the role of infrastructural instruments in promoting open research data adoption in Epidemiology, infrastructure developers need to consider enhancing various aspects of data repositories and invest in novel data infrastructures. Organizations should illustrate the benefits of open research data practices, give researchers the technical skillset needed for these practices, provide active procedural and technical support from the right supporting actors, and incentivize researchers toward these practices. As many of the findings of our case study apply to other fields, these recommendations could also be valuable for enhancing open research data practices in other research fields.

7.4. Scientific contributions of the study

Previous research has extensively examined the benefits, barriers, motivators, and factors of (open) research data sharing and reuse. Most of these studies studied open research data adoption from a general perspective, that is, without concentrating on a specific research field (Kurata et al., 2017; Sayogo & Pardo, 2013; Tenopir et al., 2011, 2015; Zuiderwijk et al., 2020). Furthermore, some publications studied the topic in specific research fields, such as Geophysics (Tenopir et al., 2018), Astrophysics (Zuiderwijk & Spiers, 2019), and Biomedicine (Piwowar & Chapman, 2010). Nevertheless, previous research has not examined the concept of open research data adoption specifically in the context of infrastructural and institutional instruments. To our best knowledge, there is only one publication (Zuiderwijk & van Gend, in press) assessing how these instruments can be used to enhance open research data adoption, but this study concentrates on a specific university (rather than on a specific research field). Therefore, the novelty of our study comes from examining open research data sharing and reuse practices by (1) using this novel concept of infrastructural and institutional instruments, and at the same time (2) performing this study in a specific field. This study is, to our best knowledge, the first study that focused on the Epidemiology field while examining the roles of instruments based on field-dependent characteristics.

Considering that previous research mostly examined open research data adoption in fields with high data sharing practices, examining the field of Epidemiology (a field in which open data sharing and reuse are considered to be at lower levels) has scientific relevance because this study enables future research to make systematic comparisons between fields with varying data sharing levels to get insights into success and failure criteria of open research data adoption. Furthermore, the conceptual framework that we established in chapter 3.3. could be a starting point for future research that will perform similar case studies that focus on other research fields. For the field of Epidemiology, the refined conceptual framework that we established in chapter 4.7.5. could be a starting point for studies that will perform similar case studies that focus on different geographies.

Finally, this research also contributes to two theories (i.e. technology acceptance models and the institutional theory) that were used in this study. With this research, we established how the prepositions of these theories can be used to understand the mechanisms by which infrastructural and institutional instruments can influence motivations toward open data practices. Our study showed the extent to which open data infrastructures can be represented as a "technology innovation" in the technology acceptance models while discussing the strengths and weaknesses of the model (chapter 4.7.6.). Finally, we showed the extent to which the multi-actor issue of open research data adoption can be represented by the social structures that are proposed by the institutional theory (chapter 4.7.6.).

7.5. Societal and managerial contributions of the study

This research essentially examines ways of addressing the barriers in front of open research data sharing and reuse. Tackling the barriers to open research data adoption benefits society in

various ways. It benefits the researchers engaging in the data sharing or reuse practice, because the researcher who shares their research data along with the publication could get more citations for the associated research article as well as citations for the research data itself (Patel, 2016). Moreover, if a researcher opens up a dataset that was used to produce a journal article, others could replicate the results and in one respect help the researcher effortlessly prove the validity of the original research results. Reusing openly shared data benefits researchers by saving them time and effort from data collection processes (Tenopir et al., 2011). Considering that Epidemiological studies could be extremely costly, this is highly valuable. Enhanced data sharing benefits academic fields by preventing research misconduct (e.g. fabrication and falsification of research data), reducing errors in research results, and building transparency to research processes (Patel, 2016; Tenopir et al., 2011). Researchers can combine multiple datasets from different sources and perform metanalyses to produce novel research findings thanks to open data practices (Institute of Medicine, 2013). Since making research datasets open increases the visibility of the researcher and research outputs, collaborations in the scientific field can also increase (Institute of Medicine, 2013). The Epidemiology field would especially benefit from increased open research data sharing since it would help understand diseases faster. This is important because accessing data is a vital prerequisite for identifying a public health problem that necessitates an urgent response (Hedberg & Maher, 2018). For instance, during the Covid-19 pandemic, by openly sharing research data about the SARS-CoV-2 genome, researchers in China helped the researchers in other parts of the world to develop critical diagnostic methods and helped facilitate pandemic response activities (Schwalbe et al., 2020).

Open research data adoption could also positively affect the public's relationship with research and researchers. Because of increased transparency of the research processes and enhanced perception of scientific knowledge being a public good, the society would build more trust in research and show more willingness to attribute funding for research. Furthermore, open research data adoption can also remove the financial barriers in front of research in low- and middle-income countries, and lower inequalities due to the imbalance of research resources across the globe (Tennant et al., 2016).

In this study, we systematically analyzed instruments to identify the faulty or lacking points about each of them, and subsequently transformed these points into specific system requirements and recommendations. This approach of transforming behavioral elements (e.g. expectations and motivations of researchers) into tangible requirements for technical and organizational environments can guide many stakeholders such as infrastructure developers/providers, university boards/policymakers, funders, and governmental policymakers, who aim at effective interventions that guarantee changes in practice. Since our study gives insights into the usability of data infrastructures in the context of open data practices, infrastructure developers can use these insights when operationalizing usability in value-based designs. Our study can inform governmental policymakers and lawmakers who want to tackle the barriers to data sharing stemming from the GDPR. Given that there are many possible strategies that can be used to increase open data sharing, university policymakers can prioritize their interventions based on our study's indications of which of these tools are more

promising than others. Funding agencies such as European Commission (EC) and NWO (Dutch Research Council) can use the results of our study to understand what type of interventions could result in increased open data sharing practices (or what kind of interventions may rather be ineffective): for example, our study indicated that simply pushing researchers to share research data in the form of (coercive) policies may be an ineffective method. Our study can help libraries to understand how they can effectively broaden their roles in research data management support and in which ways they should increase their capabilities, considering the types of support researchers (will) need in the long run. Finally, our study can show university legal teams what to consider when shaping their communications with researchers in the context of open data practices.

7.6. Suitability of the project to the CoSEM MSc program

The perspective of a study in Complex Systems Engineering and Management (CoSEM) is making a system design in a given institutional setting. Furthermore, the system that is studied is considered a complex socio-technical system. This master thesis project qualifies as CoSEM research because we used several CoSEM perspectives during evaluating the problem, designing the scope of the research, and evaluating the results of our study. First, we evaluated the issue of low open research data sharing and reuse as a socio-technical issue, because open research data adoption is not only related to technical data infrastructures in which research data sharing activities are done, but also to the researcher's motivations (behavioral aspects), to existing institutions (i.e. regulations, laws, etc.) and to external conditions; which are all strongly related to the institutional socio-environment where researchers engage in data practices. Using this perspective, we chose to examine infrastructural and institutional instruments together and in relation to one another. Moreover, before designing our case study, we evaluated the issue of open research data adoption as a complex problem, not only because there are a lot of uncertainties in how behaviors towards the phenomenon can change, but also because underlying factors behind open research data sharing and reuse could be interrelated (Kurata et al., 2017; Zuiderwijk, 2020). Using this perspective, we deducted that many of the complexities in research data sharing practices could indeed stem from contextual, field-specific characteristics, which led us to choose to perform a field-specific Epidemiology case study as a way to tackle the complexity and obtain valuable insights for the literature. Furthermore, when assessing our findings in the case study, we acknowledged the multi-level multi-actor nature of the issue to understand the factors that affect the institutional setting under which researchers engage (or do not engage) in open research data practices. For example, specifically in chapter 4.7.6., we used the institutional theory to understand the impact of different actors who put pressure on the institutional environment. Finally, throughout this study, we aimed to simultaneously address the technological component of data sharing as well as the governance issues of data sharing, which is why, for instance, our final recommendations include not only technical but also governance interventions for the system we evaluated, confirming our efforts to realize a change in the socio-technical system that we studied.

7.7. Limitations of the study

We acknowledge that there are limitations to this study concerning several points in the research approach and research methods. Firstly, in chapter 4.3., we discussed our reasoning for opting for the quota sampling approach when recruiting Epidemiology researchers (interviewees). Using this approach, we recruited interviewees from as many UMCs in the Netherlands as possible. Selecting interviewees with different characteristics (i.e. the universities they work for) indeed has advantages. However, this approach (focusing on many different organizations) also leads to a limitation: we are not able to pose any conclusions about a specific type of Epidemiology researchers (e.g. PhDs candidates or full professors) due to varying contextual aspects (i.e. the working environment). Moreover, it should be noted that the results on the roles of these instruments in this research are based on a single case, which involved interviewing ten Epidemiology researchers. Researchers who participated in this study may or may not be representative of the Epidemiology field. Therefore, this study's result cannot and should not be immediately generalized to the wider Epidemiology field without replicating and validating the study by interviewing more people in the Epidemiology field.

Another limitation of this study essentially relates to the fact that this is a qualitative study with the main data collection method being the interviews. Conducting interviews to collect data has many advantages such as being able to get in-depth information about the phenomenon that is studied. However, it also has a drawback that with this data collection method, we had to rely on the respondents' statements. We acknowledge that there is a possibility that some researchers may have given biased or unrealistic answers to our highly behavioral questions, which then would affect the validity of the results of our study. To address this, the findings that we deducted from the interviews could have been backed up by the actual observations in practice. However, because of the time limitations and scope of this master thesis, it was not possible to make such additions to our study. Another limitation concerning conducting interviews is that, arguably, it is one of the most time-consuming data collection approaches. In our study, the data collection process was the biggest challenge because we had serious issues with finding participants that would be willing to make strong time commitments.

Despite having adopted a systematic approach for operationalizing the qualitative data in this study (chapter 4.5.4.), we acknowledge that there is always the possibility of researcher bias in the data analysis procedure. We aimed to avoid bias further by giving the interviewees the chance to review the interview notes and make changes before including them in our analysis, by examining secondary sources (i.e. policy documents) to justify our deductions, and by letting third-person professionals review the results of the study and collect feedback in the form of a workshop.

7.8. Directions for future research

Our case study focused on examining how infrastructural and institutional instruments would influence the behavior of Epidemiology researchers. We recommend prospective research to conduct case studies in other contexts, considering that the issue that we examined in this

research is a multi-actor issue. Possible focus points could be examining the attitudes of policymakers in certain universities or examining the capabilities of infrastructure developers to better understand how instruments can be operationalized in practice. To address the limitation concerning our quota sampling approach, other case studies could also systematically focus on understanding behavioral differences towards open data practices among different types of researchers in an organization. Furthermore, we recommend researchers to replicate our case study with different Epidemiologists to get better insights into whether the findings of this study can be generalized to the wider population of Epidemiologists.

Moreover, we recommend future research to further examine the new instruments that emerged from our study. We recommend future research to examine whether open research data sharing motivations would be positively influenced by converging the procedures of different data repositories. We also recommend future research to examine the potential of increasing the usage of persistent identifiers (such as ORCID identifiers) in open research data adoption. Such research would be valuable to understand if lack of visibility is an essential problem in front of open research data adoption. Finally, we recommend future research to investigate to what extent the open science groups can, in practice, incentivize researchers to share or reuse open research data. Open science groups could be important for explaining the benefits and theoretical grounds of open data sharing and reuse. Currently, their interactions with Epidemiology researchers could be low, even in universities where open science groups are considered to be strong. Therefore, future research should address if and why open science groups have lower reach in certain fields and higher in others.

8. References

4TU.ResearchData. (n.d.). *Fair data refinement - what and why?* Retrieved February 28, 2022, from https://data.4tu.nl/info/en/about-your-data/fair-data-fund

About. (n.d.). Re3data.Org. Retrieved February 21, 2022, from https://www.re3data.org/about

ABOUT 4TU.RESEARCHDATA. (n.d.). 4TU.RESEARCHDATA. Retrieved February 21, 2022, from https://data.4tu.nl/info/en/about-4turesearchdata/organisation

About OpenSAFELY. (n.d.). OpenSAFELY. Retrieved June 7, 2022, from https://www.opensafely.org/about/

Ahmed, S., & Mohd Asraf, R. (2018). *The Workshop as a Qualitative Research Approach: Lessons Learnt from a "critical Thinking Through Writing' Workshop*.

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. In J. Kuhl & J. Beckmann (Eds.), *Action Control: From Cognition to Behavior* (pp. 11–39). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69746-3_2

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/https://doi.org/10.1016/0749-5978(91)90020-T

Altayar, M. S. (2018). Motivations for open data adoption: An institutional theory perspective. *Government Information Quarterly*, *35*(4), 633–643. https://doi.org/10.1016/j.giq.2018.09.006

Amnesia: High accuracy Data Anonymization. (n.d.). OpenAIRE. Retrieved February 21, 2022, from https://amnesia.openaire.eu

Barrett, D., & Noble, H. (2019). What are cohort studies? *Evidence-Based Nursing*, *22*(4), 95–96. https://doi.org/10.1136/ebnurs-2019-103183

Behnke, C., Staiger, C., Coen, G., le Franc, Y., Parland-von Essen, J., Riungu-Kalliosaari, L., & Bonino, L. (2019). *Fostering FAIR Data Practices in Europe*. https://zenodo.org/record/3631528#.YhOgvy8w01J

Bialke, M., Bahls, T., Havemann, C., Piegsa, J., Weitmann, K., Wegner, T., & Hoffmann, W. (2015). MOSAIC – A Modular Approach to Data Management in Epidemiological Studies. *Methods of Information in Medicine*, *54*(04), 364–371. https://doi.org/10.3414/ME14-01-0133

Böhmer, J. K. (2019). *Data Stewardship in the Netherlands*. GO TRAIN Workshop. https://www.go-fair.org/wp-content/uploads/2019/12/06_goTRAIN_DataStewardshipNL_V3_191125.pdf

Burgelman, J.-C., Pascu, C., Szkuta, K., von Schomberg, R., Karalopoulos, A., Repanas, K., & Schouppe, M. (2019). Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data*, *2*. https://www.frontiersin.org/article/10.3389/fdata.2019.00043

Burwell, S. M., Mancini, D. J., VanRoekel, S., & Park, T. (2013). *Memorandum for the heads of executive departments and agencies: Open data policy.* Washington D.C. https://project-open-data.cio.gov/policy-memo/

Campbell, J. (2015). Access to Scientific Data in the 21st Century: Rationale and Illustrative Usage Rights Review. *Data Science Journal*, *13*, 203–230. https://doi.org/10.2481/dsj.14-043

CDC. (2013). *Analyzing and Interpreting Large Datasets*.

Childs, S., McLeod, J., Lomas, E., & Cook, G. (2014). Opening research data: Issues and opportunities. *Records Management Journal*, *24*(2), 142–162. https://doi.org/10.1108/RMJ-01-2014-0005

Clarke, P., & Davidson, J. (2021). *Supporting the alignment of organisational research data management policies (webinar)*. https://dri.ie/rda4eosc-webinar-supporting-alignment-organisational-research-data-management-policies-7th-may-1300

Clinical Trials and Cohort Studies Grants. (n.d.). Australian Government National Health and Medical Search Council. Retrieved May 10, 2022, from https://www.nhmrc.gov.au/funding/find-funding/clinical-trials-and-cohort-studies-grants#

Creamer, A., Morales, M., Crespo, J., Kafel, D., & Martin, E. (2012). An Assessment of Needed Competencies to Promote the Data Curation and Management Librarianship of Health Sciences and Science and Technology Librarians in New England. *Journal of EScience Librarianship*, *1*, 4. https://doi.org/10.7191/jeslib.2012.1006

Crosas, M. (2016). Open Source Tools Facilitating Sharing/Protecting Privacy: Dataverse and DataTags. In *NFAIS Webinar series*. Harvard University . https://www.slideshare.net/mercecrosas/open-source-tools-facilitating-sharingprotecting-privacy-dataverse-and-datatags

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC Medical Research Methodology*, *11*(1), 100. https://doi.org/10.1186/1471-2288-11-100

Curtin, M., & Fossey, E. (2007). Appraising the trustworthiness of qualitative studies: Guidelines for occupational therapists. *Australian Occupational Therapy Journal*, *54*(2), 88–94. https://doi.org/https://doi.org/10.1111/j.1440-1630.2007.00661.x

da Costa, M. P., & Lima Leite, F. C. (2019). Factors influencing research data communication on Zika virus: a grounded theory. *Journal of Documentation*, *75*(5), 910–926. https://doi.org/10.1108/JD-05-2018-0071

Data Policy. (n.d.). OSTHUS. Retrieved February 28, 2022, from https://www.osthus.com/osthus-glossary/d/data-policy

DataverseNL: FAQ. (n.d.). DataverseNL. Retrieved May 22, 2022, from https://dataverse.nl

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: theory and results. Doctoral dissertation*.

de Knecht, S., Miedema, F., van der Meer, M., Kluijtmans, M., & Brinkman, L. (2021). *RESHAPING THE ACADEMIC SELF*. https://www.uu.nl/sites/default/files/210401%20-%20White%20Paper%20Open%20Science%20Education.pdf

Delivering Digital Infrastructure Advancing the Internet Economy. (2014). https://www3.weforum.org/docs/WEF_TC_DeliveringDigitalInfrastructure_InternetEconomy_Report_2014.pdf

Dewey, A., & Drahota, A. (2016). *Introduction to systematic reviews: online learning module Cochrane Training*. Cochrane Training. https://training.cochrane.org/interactivelearning/module-1-introduction-conducting-systematic-reviews

Dimaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. In *Source: American Sociological Review* (Vol. 48, Issue 2).

dmponline-TU Delft. (n.d.). TU Delft. Retrieved February 28, 2022, from https://dmponline.tudelft.nl/plans

Downs, R. R. (2021). Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories. *Certifying Research Data Repositories. Data Science Journal*, *20*(1), 1–11. https://doi.org/10.5334/dsj

Drucker, A. M., Fleming, P., & Chan, A.-W. (2016). Research Techniques Made Simple: Assessing Risk of Bias in Systematic Reviews. *Journal of Investigative Dermatology*, *136*(11), e109–e114. https://doi.org/https://doi.org/10.1016/j.jid.2016.08.021

European Commission. (2009). *European Commission: ICT Infrastructures for e-Science.* https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0108:FIN:EN:PDF

FAIR Data Repositories: Key Features Defined. (n.d.). FAIRSFAIR. Retrieved February 21, 2022, from https://www.fairsfair.eu/news/fair-data-repositories-key-features-defined

Fatehi, F., Gray, L. C., & Wootton, R. (2014). How to improve your PubMed/MEDLINE searches: 3. advanced searching, MeSH and My NCBI. *Journal of Telemedicine and Telecare*, *20*(2), 102–112. https://doi.org/10.1177/1357633X13519036

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, *10*(2). https://doi.org/10.1371/journal.pone.0118053

Find trustworthy data repository certified repositories. (2018, November 3). OpenAIRE. https://www.openaire.eu/find-trustworthy-data-repository-certified-repositories

Floyd, S. W. (2009). 'Borrowing' Theory: What Does This Mean and When Does It Make Sense in Management Scholarship? *Journal of Management Studies*, *46*(6), 1057–1058. https://doi.org/https://doi.org/10.1111/j.1467-6486.2009.00865.x

Gehman, J., Glaser, V., Eisenhardt, K., Gioia, D., Langley, A., & Corley, K. (2018). Finding Theory-Method Fit: A Comparison of Three Qualitative Approaches to Theory Building. *Journal of Management Inquiry*, *27*, 284–300. https://doi.org/10.1177/1056492617706029

Ghazizadeh, M., Peng, Y., Lee, J., & Boyle, L. (2012). Augmenting the Technology Acceptance Model with Trust: Commercial Drivers' Attitudes towards Monitoring and Feedback. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56). https://doi.org/10.1177/1071181312561481

Gioia, D., Corley, K., & Hamilton, A. (2013). Seeking Qualitative Rigor in Inductive Research. *Organizational Research Methods*, *16*, 15–31. https://doi.org/10.1177/1094428112452151

Gomez-Diaz, T., & Recio, T. (2022). Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context. *F1000Research*, *11*, 118. https://doi.org/10.12688/f1000research.78195.1

Gopalakrishnan, S., & Ganeshkumar, P. (2013). Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *Journal of Family Medicine and Primary Care*, *2*(1), 9–14. https://doi.org/10.4103/2249-4863.109934

Grant, C., & Osanloo, A. (2015). Understanding, selecting, and integrating a theoretical framework in dissertation research: Developing a "blueprint" for your 'house.' *Administrative Issues Journal*, *4*. https://doi.org/10.5929/2014.4.2.9

Green, S. (2005). Systematic reviews and meta-analysis. *Singapore Medical Journal*, *46(6)*, 270–274.

Ha, N., Nguyen, T., Nguyen, T., & Nguye, T. (2019). The effect of trust on consumers' online purchase intention: An integration of TAM and TPB. *Management Science Letters*, *9*, 1451–1460. https://doi.org/10.5267/j.msl.2019.5.006

Hanseth, O., & Lyytinen, K. (2010). Design Theory for Dynamic Complexity in Information Infrastructures: The Case of Building Internet. *Journal of Information Technology*, *25*. https://doi.org/10.1057/jit.2009.19

Harper, L. M., & Kim, Y. (2018). Attitudinal, normative, and resource factors affecting psychologists' intentions to adopt an open data badge: An empirical analysis. *International Journal of Information Management*, *41*, 23–32. https://doi.org/10.1016/j.ijinfomgt.2018.03.001

Harvard Dataverse. (n.d.). Harvard Dataverse. Retrieved February 21, 2022, from https://library.harvard.edu/services-tools/harvard-dataverse

Hedberg, K., & Maher, J. (2018, December). *CDC Field Epidemiology Manual: Collecting Data*. https://www.cdc.gov/eis/field-epi-manual/chapters/collecting-data.html

Henfridsson, O., & Bygstad, B. (2013). The Generative Mechanisms of Digital Infrastructure Evolution. *MIS Quarterly*, *37*(3), 907–931. http://www.jstor.org/stable/43826006

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Hodgson, G. M. (2006). What Are Institutions? *Journal of Economic Issues*, *40*(1), 1–25. http://www.jstor.org/stable/4228221

Hood, A. S. C., & Sutherland, W. J. (2021). The data-index: An author-level metric that values impactful data and incentivizes data sharing. *Ecology and Evolution*, *11*(21), 14344–14350. https://doi.org/https://doi.org/10.1002/ece3.8126

Hoofnagle, C. J., van der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98. https://doi.org/10.1080/13600834.2019.1573501

How to Make a Data Dictionary. (n.d.). OSF. Retrieved May 22, 2022, from https://help.osf.io/article/217-how-to-make-a-data-dictionary#Variable-names

Institute of Medicine. (2013). *Sharing Clinical Research Data: Workshop Summary. .* National Academies Press (US). https://doi.org/https://doi.org/10.17226/18267

Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management*, *35*(4), 408–418. https://doi.org/10.1016/j.ijinfomgt.2015.04.007

Korstjens, I., & Moser, A. (2018). Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. *European Journal of General Practice*, *24*(1), 120–124. https://doi.org/10.1080/13814788.2017.1375092

Kurata, K., Matsubayashi, M., & Mine, S. (2017). Identifying the complex position of research data and data sharing among researchers in natural science. *SAGE Open*, *7*(3). https://doi.org/10.1177/2158244017717301

LaMorte, W. W. (2021). *Epidemiologic Study Designs 1: Cohort Studies & Clinical Trials*. Boston University School of Public Health. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module4-Cohort-RCT/PH717-Module4-Cohort-RCT6.html

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR – As Open as Possible, as Closed as Necessary. *Data Intelligence*, *2*(1–2), 47–55. https://doi.org/10.1162/dint_a_00027

Last, J. M. (1988). What Is "Clinical Epidemiology?" *Journal of Public Health Policy*, *9*(2), 159–163. https://doi.org/10.2307/3343001

Last, J. M. (Ed.). (2001). *Dictionary of Epidemiology* (4th Edition). Oxford University Press.

Lin, S. (2019, July 18). *The Synthetic Data Solution: Bypassing the Headache of Data Privacy*. PLUGANDPLAY. https://www.plugandplaytechcenter.com/resources/synthetic-data-solution-bypassing-headache-data-privacy/

Linneberg, M., & Korsgaard, S. (2019). Coding qualitative data: a synthesis guiding the novice. *Qualitative Research Journal*. https://doi.org/10.1108/QRJ-12-2018-0012

Lucas-Dominguez, R., Alonso-Arroyo, A., Vidal-Infer, A., & Aleixandre-Benavent, R. (2021). The sharing of research data facing the COVID-19 pandemic. *Scientometrics*, *126*(6), 4975–4990. https://doi.org/10.1007/s11192-021-03971-6

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, *4*(3), 445–455. https://doi.org/10.1080/19439342.2012.711342

Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, *14*(1), 81–95. https://doi.org/10.1007/s10209-014-0348-1

Mcleod, S. (2019). *Case study method*. Simply Psychology. https://www.simplypsychology.org/case-study.html

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, *83(2)*, 340–363. http://www.journals.uchicago.edu/t-and-c

Michener, W. K. (2015). Ecological data sharing. In *Ecological Informatics* (Vol. 29, Issue P1, pp. 33–44). Elsevier B.V. https://doi.org/10.1016/j.ecoinf.2015.06.010

Miedema, F. (2021). *Viewpoint: As part of Global Shift, Utrecht University is changing how it evaluates its researchers*. ScienceBusiness. https://sciencebusiness.net/viewpoint/viewpoint-part-global-shift-utrecht-university-changing-how-it-evaluates-its-researchers

Miles, M., Huberman, M., & Saldaña, J. (2013). Qualitative Data Analysis: A Methods Sourcebook. In *Zeitschrift fur Personalforschung* (Vol. 28).

Mitchell, C., Brigden, T., Cook, S., & Hall, A. (2021). *Regulation and use of confidential patient information for genomic and medical research during and post COVID-19*. https://www.phgfoundation.org/media/549/download/Regulation%20and%20use%20of%20confidential%20patient%20information%20for%20genomic%20and%20medical%20research%20during%20and%20post%20COVID-19%20-%20Report.pdf?v=1&inline=1

Mohd Ishak, N., & Abu Bakar, A. Y. (2014). Developing Sampling Frame for Case Study: Challenges and Conditions. *World Journal of Education*, *4*(3). https://doi.org/10.5430/wje.v4n3p29

Motschall, E., & Falck-Ytter, Y. (2005). Searching the MEDLINE Literature Database through PubMed: A Short Guide. *Oncology Research and Treatment*, *28*(10), 517–522. https://doi.org/10.1159/000087186

Mullner, R. M. (n.d.). Epidemiology: Sources of epidemiological data. In The Editors of Encyclopaedia Britannica (Ed.), *Britannica*. Retrieved May 11, 2022, from https://www.britannica.com/science/epidemiology/Sources-of-epidemiological-data

Network Common Data Form (NetCDF). (n.d.). UNIDATA. Retrieved February 21, 2022, from https://www.unidata.ucar.edu/software/netcdf/

Neylon, C. (2017). Building a Culture of Data Sharing: Policy Design and Implementation for Research Data Management in Development Research. *Research Ideas and Outcomes*, *3*, e21773. https://doi.org/10.3897/rio.3.e21773

North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808678

OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*.

Open Science: Recognition and rewards. (n.d.). Utrecht University. Retrieved May 19, 2022, from https://www.uu.nl/en/research/open-science/tracks/recognition-and-rewards

Ørngreen, R., & Levinsen, K. (2017). Workshops as a research methodology. *Electronic Journal of E-Learning*, *15*, 70–81.
Other researchers can use your data without having to repeat the animal testing. (2021, September 30). NWO. https://www.nwo.nl/en/news/other-researchers-can-use-your-data-without-having-repeat-animal-testing
Over Lifelines. (n.d.). Lifelines Biobank Webpage. Retrieved May 10, 2022, from https://www.lifelines.nl
Parodi, B. (2015). Biobanks: A Definition. In D. Mascalzoni (Ed.), *Ethics, Law and Governance of Biobanking: National, European and International Approaches* (pp. 15–19). Springer Netherlands. https://doi.org/10.1007/978-94-017-9573-9_2
Patel, D. (2016). Research data management: a conceptual framework. *Library Review*, *65*(4/5), 226–241. https://doi.org/10.1108/LR-01-2016-0001
Pearce, N., & Smith, A. H. (2011). Data sharing: Not as simple as it seems. In *Environmental Health: A Global Access Science Source* (Vol. 10, Issue 1). https://doi.org/10.1186/1476-069X-10-107
Pittway, L. (2008). *The SAGE dictionary of qualitative management research* (R. Thorpe & R. Holt, Eds.). SAGE Publications.
Piwowar, H. A., Becich, M. J., Bilofsky, H., & Crowley, R. S. (2008). Towards a data sharing culture: Recommendations for leadership from academic health centers. In *PLoS Medicine* (Vol. 5, Issue 9, pp. 1315–1319). https://doi.org/10.1371/journal.pmed.0050183
Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics*, *4*(2), 148–156. https://doi.org/https://doi.org/10.1016/j.joi.2009.11.010
Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, *2*(3), e308-. https://doi.org/10.1371/journal.pone.0000308
Poulis, G., Loukides, G., Skiadopoulos, S., & Gkoulalas-Divanis, A. (2017). Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints. *Journal of Biomedical Informatics*, *65*, 76–96. https://doi.org/https://doi.org/10.1016/j.jbi.2016.11.001
Principles of Epidemiology in Public Health Practice: An introduction to applied epidemiology and biostatistics. (2012). U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. https://www.cdc.gov/csels/dsepd/ss1978/index.html
Qualitative Data Analysis: The Research Guide. (2020). https://www.nhgeducation.nhg.com.sg/homer/Pages/Research/Research-Guides1120-1511.aspx
Rao, S., & Reddy, V. (2013). AN EXAMINATION OF THE ROLE OF CONCEPTUALIZATION AND OPERATIONALIZATION IN EMPIRICAL SOCIAL RESEARCH. *International Journal of Multidisciplinary Research*, *3(7)*.
Recognition & Rewards - Research. (n.d.).
Reeves, S., Albert, M., Kuper, A., & Hodges, B. (2008). Qualitative research - Why use theories in qualitative research? *BMJ (Clinical Research Ed.)*, *337*, a949. https://doi.org/10.1136/bmj.a949
Research data management support: Policies, codes of conduct and laws. (n.d.). Utrecht University. Retrieved May 20, 2022, from https://www.uu.nl/en/research/research-data-management/guides/policies-codes-of-conduct-and-laws#ownership
Restricting access to data. (n.d.). University of York. Retrieved May 10, 2022, from https://www.york.ac.uk/library/info-for/researchers/data/sharing/access/

Reuse of already collected datasets (secondary use). (n.d.). Tiburg University. Retrieved May 10, 2022, from https://www.tilburguniversity.edu/about/conduct-and-integrity/privacy-and-security/research-data/datasets/reuse

Ringersma, J., & Adamse, P. (2019). *Data Stewardship @ WUR: advice on a role for Data Stewards*. https://zenodo.org/record/2561723

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*(SUPPL. 1). https://doi.org/10.1016/j.giq.2012.06.011

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLOS ONE*, *11*(1), e0146695-. https://doi.org/10.1371/journal.pone.0146695

Schwalbe, N., Wahl, B., Song, J., & Lehtimaki, S. (2020). Data Sharing and Global Public Health: Defining What We Mean by Data. *Frontiers in Digital Health*, *2*, 612339. https://doi.org/10.3389/fdgth.2020.612339

Scott, W. (2005). *Institutional Theory: Contributing to a Theoretical Research Program*.

Scott, W. (2013). *INSTITUTIONS AND ORGANIZATIONS: Vol. 4th edition*.

Shamsuddin, A., Sheikh, A., & Keers, R. N. (2021). Conducting Research Using Online Workshops During COVID-19: Lessons for and Beyond the Pandemic. *International Journal of Qualitative Methods*, *20*, 16094069211043744. https://doi.org/10.1177/16094069211043744

Sharp, J. H. (2006). *Development, Extension, and Application: A Review of the Technology Acceptance Model*.

Shelly, M., & Jackson, M. (2018). Research data management compliance: is there a bigger role for university libraries? *Journal of the Australian Library and Information Association*, *67*(4), 394–410. https://doi.org/10.1080/24750158.2018.1536690

Strategy for a European Data Infrastructure. (n.d.). European Commissione Europa. Retrieved February 21, 2022, from https://ec.europa.eu/eurostat/cros/content/36-strategy-european-data-infrastructure_en

Sullivan, C., & Burger, E. (2017). "In the public interest": The privacy implications of international business-to-business sharing of cyber-threat intelligence. *Computer Law & Security Review*, *33*(1), 14–29. https://doi.org/https://doi.org/10.1016/j.clsr.2016.11.015

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of Open Access: An evidence-based review. *F1000Research*, *5*. https://doi.org/10.12688/f1000research.8460.1

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Birch, B., & Allard, S. (2012). *Current Practices and Plans for the Future*.

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. https://doi.org/10.1029/2018EA000461

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

The FAIR data principles. (n.d.). FORCE11. Retrieved February 21, 2022, from https://force11.org/info/the-fair-data-principles/

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Digital Infrastructures: The Missing IS Research Agenda. *Information Systems Research*, *21*, 748–759. https://doi.org/10.1287/isre.1100.0318

TU Delf Recognition & Rewards Perspective (2021-2024). (n.d.). Retrieved May 23, 2022, from https://d2k0ddhflgrk1i.cloudfront.net/TUDelft/Over_TU_Delft/Strategie/Erkennen%20en%20Waarderen%20-%20recognition%20and%20reward/TU%20Delft%20Recognition%20and%20Rewards%20perspective%20def.pdf

University of Colorado. (n.d.). *Data Management Guide*. Retrieved February 28, 2022, from https://guides.library.csupueblo.edu/data/repositories

Utrecht University. (n.d.). *Experienced Data Managers*. Retrieved February 28, 2022, from https://www.uu.nl/en/research/research-data-management/tools-services/experienced-data-managers

van Roode, M., Dos, C., Ribeiro, S., Farag, E., Ahmed, M., Moustafa, A., van de Burgwal, L., Claassen, E., Nour, M., Haringhuizen, G., & Koopmans, M. (2018). *The case of Middle East Respiratory Syndrome (MERS)*.

Venkatesh, V., & Davis, F. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*, 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., & Davis, F. D. (1996). *A Model of the Antecedents of Perceived Ease of Use: Development and Test**.

Vlahou, A., Hallinan, D., Apweiler, R., Argiles, A., Beige, J., Benigni, A., Bischoff, R., Black, P. C., Boehm, F., Céraline, J., Chrousos, G. P., Delles, C., Evenepoel, P., Fridolin, I., Glorieux, G., van Gool, A. J., Heidegger, I., Ioannidis, J. P. A., Jankowski, J., … Vanholder, R. (2021). Data Sharing Under the General Data Protection Regulation. *Hypertension*, *77*(4), 1029–1035. https://doi.org/10.1161/HYPERTENSIONAHA.120.16340

Waiting lists still increasing at hospitals, clinics. (2020, January 17). *NLTimes*. https://nltimes.nl/2020/01/17/waiting-lists-still-increasing-hospitals-clinics

What is a Data Management Plan? (n.d.). Wageningen University and Research. Retrieved February 28, 2022, from https://www.wur.nl/en/show/What-is-a-Data-Management-Plan.htm

What is a metadata standard? (n.d.). University of Pittsburgh Library System. Retrieved February 21, 2022, from https://pitt.libguides.com/metadatadiscovery/metadata-standards

What is data sharing? (n.d.). Support Centre for Data Sharing. Retrieved March 8, 2022, from https://eudatasharing.eu/what-data-sharing

What is research data? (n.d.). University of Leeds Library. Retrieved March 8, 2022, from https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

WILEY. (n.d.). *Wiley's Data Citation Policy*. Retrieved February 28, 2022, from https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-citation-policy.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Williamson, C. (2009). Informal institutions rule: Institutional arrangements and economic performance. *Public Choice*, *139*, 371–387. https://doi.org/10.1007/s11127-009-9399-x

Wirth, F. N., Meurers, T., Johns, M., & Prasser, F. (2021). Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01602-x

Yin, R. K. (2018). *Case Study Research and Applications.* (6th Edition). SAGE.

Zhang, A. B., & Gourley, D. (2009). 4 - Metadata strategy. In A. B. Zhang & D. Gourley (Eds.), *Creating Digital Collections* (pp. 31–53). Chandos Publishing. https://doi.org/https://doi.org/10.1016/B978-1-84334-396-7.50004-3

Zuiderwijk A. (2015). Open data infrastructures : the design of an infrastructure to enhance the coordination of open data use. Uitgeverij BOXPress.

Zuiderwijk, A. (2020). Open Research Data sharing and use by means of infrastructural and institutional arrangements [Keynote Presentation]. In *International Conference on ICT enhanced Social Sciences and Humanities*. https://www.youtube.com/watch?v=Mz0r0GaCh6A

Zuiderwijk, A., Shinde, R., & Jeng, W. (2020). What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. *PLoS ONE*, *15*(9 September). https://doi.org/10.1371/journal.pone.0239283

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. https://doi.org/10.1016/j.ijinfomgt.2019.05.024

Zuiderwijk, A., & van Gend, T. (in press). Open research data: a case study into institutional and infrastructural arrangements to stimulate open research data sharing and reuse. https://doi.org/10.1177/09610006221101200

9. Appendices

Appendix A: Email template for recruiting participants

My name is Berkay Türk, I am a second year Complex Systems Engineering and Management MSc student at TU Delft. I am currently conducting my master thesis project, which is a case study in the field of epidemiology on open research data sharing and reuse.

I examine the infrastructural and institutional instruments that stimulate openly sharing and reusing research data in the field of epidemiology. My objective is to understand what role such instruments can play in promoting open data sharing and use behaviour in epidemiology. Such instruments can potentially address barriers in front of open research data sharing and reuse, which is why it is valuable for me to understand whether you have access to such instruments and how they may influence your behaviour.

I have identified you as a researcher working in the field of epidemiology. I would be very interested in speaking to you about your experiences, thoughts and attitudes towards open research data sharing and reuse in your field. Would you be available for a one-hour interview?

As a sign of appreciation, I will provide you with the outcomes of this study before they will be published and give you the opportunity to comment on the findings.

I'm looking forward to your response.

With kind regards,
Berkay Türk

Appendix B: Interview design (for Epidemiology researchers)

Research Data Sharing and Reuse: a Case Study in Epidemiology – Interview Questions

Interview information
Name interviewer: Berkay Türk
Interview number:
Interview date:

Title and name respondent:
Organization where the respondent words:
Country where the respondent works: Netherlands

Introduction
Welcome
Welcome, and thank you for participating in this research. My name is Berkay Türk, I am a second-year master student at TU Delft. I conduct this research as part of my master thesis project, which is a case study in the field of Epidemiology on open research data sharing and reuse.

Research objective
In this study, we examine the infrastructural and institutional instruments that are used in the field of Epidemiology to stimulate openly sharing and reusing research data. This study's objective is to understand what role infrastructural and institutional arrangements can play in promoting open data sharing and use behaviour in Epidemiology.

I will ask you about 16 questions in five categories.
The interview takes 1 hour.
I will share my notes with you for you to revise and approve before including them in my thesis.

After the interview
When I conclude my master thesis project in July or August this year, I will share my master thesis project report with you.

Interview questions

This interview consists of five sections, namely:
1. Background information
2. Your involvement in open research data, open research data sharing, and reuse
3. Infrastructural instruments that influence your motivation and behaviour towards openly sharing and re-using research data
4. Institutional instruments that influence your motivation and behaviour towards openly sharing and re-using research data
5. Barriers to open research data sharing and reuse

Section 1: Background information

In this section, you are asked to provide information about your background. This data is personal and we will only report on this anonymously.

1. What is your age?

2. What is your current position in your institution?

- ○ Ph.D. candidate
- ○ Postdoctoral researcher
- ○ Assistant professor
- ○ Associate professor
- ○ Full professor
- ○ Other (please specify)

_____

3. For how long (i.e. years) have you been in this academic/scientific position?

4. In which subfield of Epidemiology are you currently employed?

Section 2: Experience with open research data sharing and reuse

My research focuses on open research data. When I refer to open research data, I mean data that is structured, machine-readable data, actively published on the internet. Open data is published for public reuse, and is ideally also Findable, Accessible, Interoperable, and Reusable (FAIR). Open data can be both quantitative and qualitative and can be either raw/primary, derived from primary data for subsequent analysis or interpretation, or derived from existing sources.

5. Do you have any experience with openly sharing research data? Could you provide more detail?
6. Do you have any experience with reusing research data that others have openly shared? Could you provide more detail?

Section 3: Infrastructural instruments for openly sharing and re-using research data

In this study, we examine the infrastructural and institutional instruments that are used to stimulate openly sharing and re-using open research data.

I would now like to discuss the infrastructural instruments with you, to understand to what extent these instruments affect your research data sharing and reuse behaviour.

In my literature review on infrastructural instruments, I have identified various infrastructural instruments that may affect open data sharing and reuse motivation and behaviour. Please note that when I refer to '(open data) infrastructures' in the following questions, I mean (technical)

infrastructures that you can use when you are engaging with open data sharing and reuse activities. Such infrastructures could be data repositories; software and tools that can be used for finding, storing, curating data, metadata creation, anonymization, analysis, research data management, licensing, etc.

7. Could you indicate, for each of the following instruments, whether you have access to these instruments and explain to what extent they affect your open research data sharing and reuse activities? Please answer the questions according to the infrastructures that you use.

Infrastructural instruments	**Do you agree with the statement? (yes, no, partially) Please explain why.**	**To what extent does this instrument influence your open research data sharing and reuse behaviour?**
The open data infrastructure(s) that I use has (have) user-friendly graphic interfaces.		
The open data infrastructure(s) that I use allows (allow) for easy and quick data analysis.		
The open data infrastructure(s) that I use is (are) easy to use.		
The open data infrastructure(s) that I use is (are) compatible with the different types of data that are used in my field.		
The open data infrastructure(s) that I use is (are) reliable.		
The open data infrastructure(s) that I use can accommodate a large volume of data.		
The open data infrastructure(s) that I use helps (help) with choosing a license (e.g. CC0, CC-BY., etc.).		
The infrastructures (e.g. data repositories; software that are used for metadata creation, anonymization, etc.) that I use are integrated with each other and compatible.		
The search engine(s) on the open data repository that I use is (are) sufficient for my open data search needs.		
The open data infrastructure(s) that I use helps (help) with the		

Infrastructural instruments	Do you agree with the statement? (yes, no, partially) Please explain why.	To what extent does this instrument influence your open research data sharing and reuse behaviour?
selection of an appropriate repository for openly sharing research data.		
A data management tool is offered to me.		
The open data repository that I use accommodates metadata standards (i.e. enabling to properly store metadata and to view metadata of other datasets)		
The open data repository that I use helps with the creation of an appropriate citation for the data.		
Tools for metadata creation and management are offered to me.		
The open data infrastructure(s) that I use incentivizes (incentivize) usage of metadata standards (e.g. by explicitly asking for the usage of standards when sharing your research data).		
The open data repository that I use is linked to overarching/ aggregating infrastructures (i.e. registry of repositories) which help searching for data across different data repositories.		
The open data repository that I use requires the data depositor to provide metadata on the data collection methods.		
The open data infrastructure(s) that I use presents (present) data usage statistics.		
The open data infrastructure(s) that I use is (are) trustworthy (e.g. in terms of securely storing data, against breach).		
Ways/methods to assess how trustworthy an open data repository or an open data set are offered to me.		
Tools for data anonymization are offered to me.		

Infrastructural instruments	Do you agree with the statement? (yes, no, partially) Please explain why.	To what extent does this instrument influence your open research data sharing and reuse behaviour?
On the open data repository that I use, there are different access restriction types to choose from (i.e. giving ability to place conditions on data access).		

8. Are there any other existing infrastructural instruments that make it easier for you to participate in open data sharing and reuse, or that incentivize or facilitate open data sharing and reuse?

9. Can you think of any other functionalities that you wish open data infrastructures had, so that you would be more stimulated towards openly sharing and reusing research data? (Alternatively, you can indicate the troublesome features about the infrastructures that you wish would be fixed)

10. Which infrastructural instruments do you believe stimulate openly sharing research data and reusing openly shared research data the most in your field?

Section 4: Institutional instruments for openly sharing and re-using research data

I would now like to discuss institutional instruments for open research data. When I refer to *institutional instruments,* I refer to the combination of formal structures (e.g., policy, processes), informal structures (e.g. norms, culture), and more enforcing or operational mechanisms that institutions can implement to stimulate openly sharing research data and reusing open research data.

11. Please answer the questions according to the institutional context under which you conduct research and engage in open research data sharing and reuse activities.

Institutional instruments	Do you agree with the statement? (yes, no, partially) Please explain why.	To what extent does this instrument influence your open research data sharing and reuse behaviour?
There is an institutional data sharing policy and/or there are guidelines for openly sharing or reusing research data in my organization.		
There is an institutional data management policy and/or a data		

Institutional instruments	Do you agree with the statement? (yes, no, partially) Please explain why.	To what extent does this instrument influence your open research data sharing and reuse behaviour?
deletion policy and/or data security policy in my organization.		
My organisation requires me to create a data management plan (DMP) as a necessary part of the research cycle. (i.e. asking the researcher to think about costs related to access, management, and preservation of data before the research starts)		
My organisation provides support for understanding and fulfilling legal requirements ("legal basis for rights of use") regarding openly sharing or reusing research data.		
Different data management policies and guidelines (that I am aware of) are aligned and consistent with one another.		
My organisation provides guidelines on obtaining consent for (open) data sharing.		
My organisation provides guidelines on data anonymization.		
My organization's library provides support regarding non-technical topics such as choosing appropriate open data tools, selection of repositories; and/or regarding technical support such as digital curation of data, preparing datasets for a repository, accessing a repository, archiving data, backup practices, removing data from repositories, and creating metadata for datasets.		
My organization's legal departments provide support regarding open research data sharing and reuse. (Such support could be in terms of privacy, data ownership, copyright, etc.)		
My organization's data stewards provide support. (e.g., possibility to ask questions to the data steward		

Institutional instruments	Do you agree with the statement? (yes, no, partially) Please explain why.	To what extent does this instrument influence your open research data sharing and reuse behaviour?
about open research data sharing and reuse)		
I have the ability to hire data managers to take care of my research data management activities.		
My organization's website gives information and guidance on data management and open data sharing and reuse requirements. (e.g. researcher can easily reach relevant guidelines via organization's websites)		
My organization provides training and educational support (seminars, courses, training modules, etc.), for instance, on topics of open science, data management, technical training on archiving/backup, digital description or curation of data sets, data anonymization, etc.		
My organization or my field offers financial resources such as separate funds for treatment and management of openly shared research data.		
(open) data sharing is framed as a concrete goal in my organization and my organization's policy (documents).		
There is a data-sharing culture in my organization.		
Open research data contributions are recognized and rewarded in my organization and/or in my field.		
Data sharing contributions are considered during hiring, tenure, and/or promotion decisions in my organization or my field.		
My organization and/or my field uses track metrics for data sharing contributions.		
There is a data citation policy in my organization.		

Institutional instruments	**Do you agree with the statement? (yes, no, partially) Please explain why.**	**To what extent does this instrument influence your open research data sharing and reuse behaviour?**
The academic journals in my field mandate or request me to openly share research data.		
The funders in my field mandate or request me to openly share research data.		
My organization mandates or requests me to openly share research data.		
My organization publishes experiences in research data sharing on its website to promote open data.		
My organization helps me to comprehend the benefits of data sharing and the needs for data sharing.		
My organization helps me to understand ways to tackle issues around data ownership, ethics, and privacy in open data.		

12. Are there any other existing institutional instruments that facilitate or incentivize open data sharing and reuse? (You can think of such instruments by considering what kind of formal (policies, rules), informal structures (norm, culture), or enforcement mechanisms are used in your organization that incentivize, ease or facilitate your open research data sharing and reusing behaviour.)

13. Apart from the topics we discussed, what kind of support do you wish to receive from your organisation so that you would be more stimulated towards openly sharing and reusing research data? (Alternatively, you can indicate the troublesome institutional/organizational issues that that you wish would be solved)

14. Which institutional instruments do you believe stimulate openly sharing research data and reusing openly shared research data the most in your field?

Part 5: Barriers to open research data sharing and reuse

15. Despite the discussed instruments, which factors inhibit *openly sharing* research data on a large scale in your field?

16. Despite the discussed instruments, which factors inhibit the *reuse* of open research data on a large scale in your field?

Finalizing the interview

This is the end of this interview. Thank you very much for your cooperation. I will process this interview and send you a summary of your answers to the questions so that you will be able to review your answers and/or add anything you wish.

Do you have any final questions or additions concerning this interview?
…….

Then I wish you a nice day and again thank you so much for your participation in my research.

Appendix C: Interview design (for the research data management consultant)

Research Data Sharing and Reuse: a Case Study in Epidemiology – Interview Questions

Introduction

Research objective

In this study, we examine the infrastructural and institutional instruments that are used in the field of Epidemiology to stimulate openly sharing and reusing research data. This study's objective is to understand what role infrastructural and institutional arrangements can play in promoting open data sharing and use behaviour in Epidemiology. So far, my primary source of information has been epidemiology researchers.

During the interviews I had with Epidemiology researchers, it was brought up by several researchers that the legal context forming the boundary of data sharing in Epidemiology influences open data practices in the field heavily. It was mentioned several times that open research data for the field of human health is usually not fully guided by individual researchers' behaviour, but rather it is significantly bounded by GDPR privacy laws and informed consent procedures. Therefore, in light of these developments I wanted to interview somebody who has expertise on data protection issues to fully understand how the issues stemming from privacy regulations are, and to what extent these can be tackled using infrastructural and institutional instruments.

I will ask you about 20 questions in 4 categories.
The interview takes roughly 40 mins
I will share my notes with you for you to revise and approve before including them in my thesis.

After the interview

When I conclude my master thesis project in July or August this year, I will share my master thesis project report with you.

Interview questions

This interview consists of five sections, namely:

1. Background information
2. Open research data sharing and reuse in the field of Epidemiology
3. GDPR and informed consent as barriers to open research data sharing
4. Infrastructural and institutional instruments that are used to stimulate openly sharing and reusing research data

Section 1: Background information

In this section, you are asked to provide information about your background. This data is personal and we will only report on this anonymously.

1. What is your current position in your institution?
2. Do you have a specific research theme (field) that you are engaging with as part of your current position?

Section 2: Open research data sharing and reuse in the field of Epidemiology

My research focuses on open research data. When I refer to open research data, I mean data that is structured, actively published on the internet. Open data is published for public reuse, and is ideally also Findable, Accessible, Interoperable, and Reusable (FAIR). Open data can be both quantitative and qualitative and can be either raw/primary, derived from primary data for subsequent analysis or interpretation, or derived from existing sources.

3. How does your current position relate to open research data sharing in epidemiology? How do you engage/work with epidemiology researchers regarding open research data sharing?
4. Can you talk about the characteristics of open data sharing practices in the field of epidemiology?

-Which repositories exist for epidemiology?
-What type of data types are being openly shared in the field of epidemiology? What are the characteristics of the data that are being shared openly?
-What is, in your opinion, the level of open research data sharing in the field?
-How much influence do researchers actually have on open data sharing or reuse?
-Which types of data sharing are more prevalent in the field? (open data sharing, data sharing by collaboration, data sharing by request, etc)
-How do data sharing practices differ from other close or far research fields? Does the level of data sharing in epidemiology differ with respect to other research fields?

5. What are the biggest barriers to open data sharing and reuse activities in the field of epidemiology? What are the biggest struggles for researchers in your opinion? What demotivates them?

Section 3: GDPR and informed consent as barriers to open research data sharing

6. What are the current GDPR and informed consent regulations that affect open research data sharing and reuse practices in the field of Epidemiology? What do researchers have to do in order to comply with the requirements if they would like to openly share research data or reuse openly shared research data?
7. How do GDPR and informed consent requirements affect (the level of) open research data sharing in the field of epidemiology? Do you believe that these regulations and requirements are important barriers to open data practices?
8. Are there any other regulations/requirements that affect (or come as a barrier to) open research data sharing and reuse practices in the field of epidemiology?
9. Do you have any ideas on how the barriers to open data practices due to GDPR and informed consent regulations could be solved?
10. Apart from sharing primary/raw data, what are the possibilities of making metadata or summary statistics openly available in the field of epidemiology? Do GDPR and informed consent regulations also pose a problem to sharing such types of data?

Section 4: Infrastructural and institutional instruments that are used to stimulate openly sharing and reusing research data

11. Does your organization have any institutional (open) data sharing policies or policies for research data management? How do these policies affect open data sharing and reuse practices?

12. Do you think your organization provides sufficient support for explaining the legal requirements of open data practices to researchers and helping them comply with such requirements?

13. How does your organization try to promote/facilitate open data sharing and reuse practices? Do you think these promotion activities are sufficient for the field of epidemiology?

14. Do you think offering more financial resources such as separate funds for treatment and management of openly shared research data could be a positive influence for open data practices in the field of epidemiology?

15. What kind of support do you think researchers are expecting to receive from your organisation so that they would be more stimulated towards openly sharing and reusing research data? What are the troublesome organizational issues with which the researchers are struggling regarding open data practices?

16. Does your institution have any data infrastructure (such as an open data repository, or a specific data (management) software) that has been built or adopted to promote open data practices?

17. To what extent would data anonymization address the issues on open research data sharing in the field of epidemiology? What are the roles, functions, benefits of data anonymization tools in this regard?

18. Do you think the ease of use of data repositories is important for open data practices?

19. Do you think the availability of a search engine that satisfies open data search needs is important for open data practices?

20. Are there any functionalities/features that researchers wish to see in the_open data infrastructures (such as the open data repositories), so that they would be more stimulated towards openly sharing and reusing research data? What are the troublesome features of open data infrastructures with which the researchers are struggling regarding open data practices?