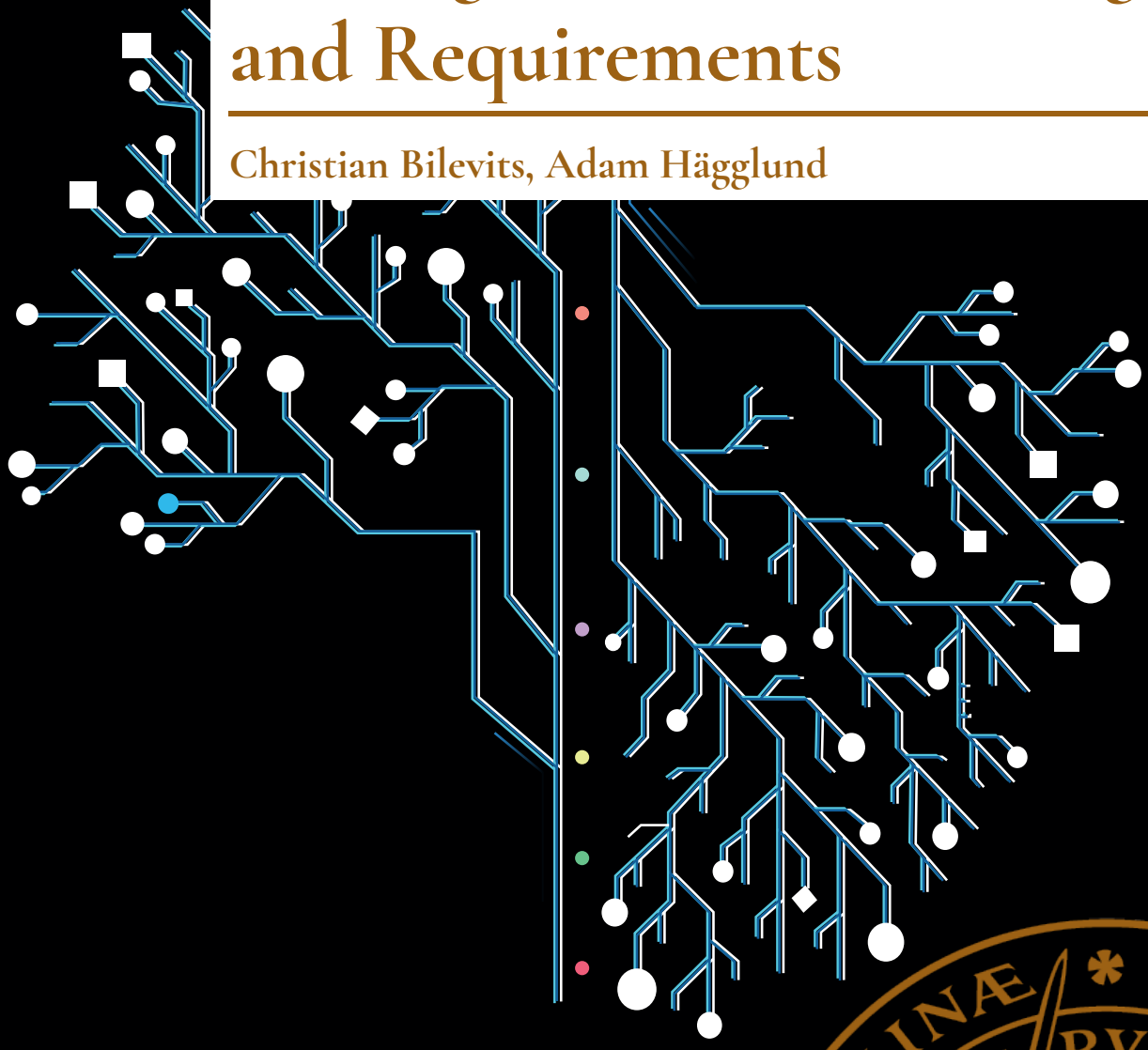
MASTER'S THESIS 2022

Data Ecosystem as a Solution for Intra-Organizational Data Sharing: Benefits, Challenges and Requirements

Christian Bilevits, Adam Hägglund



ISSN 1650-2884
LU-CS-EX: 2022-06

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2022-06

Data Ecosystem as a Solution for Intra-Organizational Data Sharing: Benefits, Challenges and Requirements

Dataekosystem som en lösning för datadelning inom ett företag: fördelar, utmaningar och krav

Christian Bilevits, Adam Hägglund

Data Ecosystem as a Solution for Intra-Organizational Data Sharing: Benefits, Challenges and Requirements

(A Master Thesis)

Christian Bilevits
christian.bilevits.8466@student.lu.se

Adam Hägglund
adam.hagglund.1116@student.lu.se

March 1, 2022

Master's thesis work carried out at Axis Communications AB.

Supervisors: Per Runeson, per.runeson@cs.lth.se
Johan Linåker, johan.linaker@cs.lth.se
Anton Friberg, anton.friberg@axis.com

Examiner: Martin Höst, martin.host@cs.lth.se

Abstract

The industrial landscape is in the midst of a rapid transformation toward becoming more Data-driven (e.g., increased use of machine learning, making decisions, and fostering innovation), and Axis Communication, having a portfolio of digital network products, is no different. Due to an increased interest in data existing within Axis, the department of Diagnostics and Data Management wants to investigate how to enable increased discovery and more efficient collaboration around data.

Correspondingly in academia, the concept of a socio-technical network for interacting and collaborating around data under the name of Data Ecosystems has started to emerge.

This master's thesis explores the benefits and challenges of data sharing in Data Ecosystems when applied in an intra-organizational context, further establishing a set of requirements towards a technical platform to address the findings. Our research follows the design science paradigm, consisting of a literature study, case study, and qualitative synthesis.

Our research concludes that although a novel concept, the current benefits, and challenges discovered at the case company correspond well with those found in the literature of Data Ecosystems. Based on the findings, 14 requirements toward a technical platform were derived and evaluated.

Keywords: Data, Metadata, Data Ecosystem, Intra-organizational, Platform-centric, Data Catalogs, Industrial Case Study

Acknowledgements

We would like to express our gratitude to our supervisors Per Runeson, Johan Linåker, and Anton Friberg for their valuable guidance and ideas throughout this thesis. We would also like to thank our examiner Martin Höst, the employees at Axis Communications, and the other master's thesis students at the department. Furthermore, we would especially like to express gratitude towards the employees at Axis who took their time to participate in our interviews. Lastly we would like to thank Mathias Bilevits[1] for designing our cover.

[1]https://www.mathiasbilevits.com

Contents

Chapter 1

Introduction

In the age of digital transformation, data has grown to be a valuable resource where its value has gained a central role and is recognized by various entities within our modern society [9]. Even though this is the case, recent research also indicates that the value of data has not yet reached its full potential, nor is it evenly distributed [9].

Data within large organizations can easily end up hidden in silos as it is collected. As an effect, finding and acquiring the data located and stored under the different departmental data silos can become a relatively complex and tedious process to perform [18]. As many projects grow toward a more data-intense workflow, the loss of potential value due to friction connected to finding and accessing data is therefore an issue they want to mitigate.

Sharing data is a topic of increasing interest in both the industry and research. An emerging concept relating to this is introducing a socio-technical Data Ecosystem, where the ecosystem aims to serve and allow actors to interact and collaborate on data with each other [30] - whether it be companies, departments, or individuals. However, infrastructure, governance, license, and privacy concerns are only a few examples of challenges that have to be considered when it comes to data sharing in a collaborative ecosystem.

One crucial pillar to potentially address the challenges mentioned is connected to the technical platform that would serve such an ecosystem. The platform could potentially enable the actors to handle the nuanced list of challenges and issues that can be present for data. Also expressed by recent literature for the evolving nature of Data Ecosystems [32].

1.1 Problem Statement

Axis Communications AB is a Swedish company founded in 1984 in Lund which provides network solutions for security purposes. As of October 2021, Axis employs 3800 people in more than 50 countries and have their headquarters located in Lund, Sweden [6]. Currently, they develop and deploy network-connected products in the areas of video surveillance, access control, and audio solutions where they distribute and sell all of their products through resellers, and system integrators, forming a global network of partners [6].

Modern businesses are becoming more data-driven, where their data assets are recognized as key components for innovation [18]. Axis, having a portfolio of digital network products, is no different.

The department Diagnostics and Data Management (DDM) focuses on collecting, storing, and handling device diagnostics telemetry data. As maintainers of data assets, DDM has received multiple requests for different kinds of data, by both employees and the management teams within the organization. For example, Quality Assurance (QA) asked for temperature distribution for different products or platform management wanted to know which firmware has the lowest risk of crashing given a population. Given this increased interest from other actors within the company to work and collaborate on their data, they are motivated to research the topic of how to better discover, share and collaborate on data.

In academia Data Ecosystems, e.g., Open Data Ecosystems (ODE) manifesting ideas and principles of Open Innovation (OI) and Open-Source Software (OSS), has recently had an increase of attention in research [30, 35]. As principles and values from OSS have been incorporated within organizations through Inner-Source [8], the same can be imagined to be done for Open Data Ecosystems into a private context, i.e., Private Data Ecosystems.

This master's thesis aims to research what benefits and challenges exist and appear when introducing a Data Ecosystem in an intra-organizational context, in order to enable data sharing, discovery, and related activities. DDM were also interested in if any platform could support these ideas. To conduct our research endeavor and gather the data for investigation, following the design science paradigm a literature study exploring the existing academic research around the research topic, in correlation with a case study at Axis Communication, was performed.

The second goal of our research was to investigate what requirements should be put on a technical platform to address and mitigate challenges and issues found.

1.2 Research Questions

Data Ecosystems being a relatively new and emerging concept in academia, in combination with the interest at the case company Axis, we were interested in further investigating the benefits and challenges of such a solution for data sharing in the existing literature.

As the problem at Axis resides in the private setting of the company, i.e., intra-organizational context, we were further interested in investigating the benefits and challenges explicitly appearing when introducing the concept of Data Ecosystem in this context.

Finally, as the department of DDM was interested in the use of a technical platform to better discover, share and collaborate on their data, a final area of interest was to investigate what requirements towards a technical platform our findings would propose.

Given our goals, the following research questions were formed:

RQ 1 What is the state-of-the-art research regarding challenges and benefits with data sharing in Data Ecosystems?
RQ 2 What are the challenges and benefits with data sharing at the case company if introducing a Data Ecosystem for private data?
RQ 3 What requirements should a technical platform fulfill to address the challenges found in RQ1 and RQ2?

1.3 Contribution

This thesis work aims to provide insight into the benefits and challenges that exists in the literature, further appearing when the emerging concept of Data Ecosystem is introduced in an intra-organizational context. Our contribution resulted in an extension of the body of knowledge around Data Ecosystems by applying an existing conceptual model in a private setting, further populating it with our empirical findings. Another aim was to investigate what requirements can be put on a technical platform to support such an ecosystem, which resulted in a set of 14 requirements further validated in the context of the problem.

Our research comprises one design cycle of the activities' problem conceptualization, solution design, and validation following the design science paradigm. To understand and build our theoretical and practical knowledge of the problem, and later support the design solution, we conducted a literature study followed by a case study at Axis Communication. As for solution design, we then synthesized and formalized a set of requirements towards a technical platform from our previously recorded material, which finally were empirically validated.

Hopefully, this research will provide new findings, insight, and guidelines for introducing a Data Ecosystem in an industry context, i.e., a Private Data Ecosystem. The thesis also aims to provide a set of technical requirements towards a platform, as well as an initial analysis of a couple of data catalog platforms to support such an ecosystem.

1.4 Division of Work

Both authors have collaborated in all activities throughout this thesis. This was partly done by using Github[1] and making pull requests[2] for the other person to review and give their opinion. However, throughout the thesis, Adam focused more on reporting the literature study and Christian the case study.

[1]https://github.com

[2]https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-pull-requests

Chapter 2

Background

This chapter aims to provide a summary of the relevant theory and related work for this master's thesis, and is divided into the sections of; Data, Data Ecosystems and Data Catalog.

As our research originates from and further aims contribute to the research field of Data Ecosystems, the first two sections; Data and Data Ecosystems provides a presentation of the theory which helped us to understand, relate and extend our research onto.

The following and final chapter presents the concept of Data Catalogs, and existing implementations of such, which in our research is set to scope our focus of existing technological platforms later used in our validation of our research.

2.1 Data

Data is a valuable resource and growing fast, or as stated in 2006 by Humby, *"data is the new oil"* [5]. According to the Cambridge Dictionary, *data* is defined as *"information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer"* [12]. Data is a general word with many meanings and different values depending on its context and characteristics.

2.1.1 Characteristics and the Value of Data

There are different types of data depending on their origin and context. Runeson et al. [35] present a categorization into seven broad categories; Maps, Society, Position, Images, Sensors, Human and Business.

- *Map* data is all kind of map data of physical and/or non-physical places.

- *Society* data is all data related to society, this could partly include map data but with added information about infrastructure, buildings, electricity etc.
- *Position* data, this is also partly related to map data, but is more focused on movements.
- *Images* data is all types of images, of faces, plants etc.
- *Sensor* data is all measurement data; light, temperature, humidity etc.
- *Human* data is data connected to humans which could be e.g. health or behavior. This type of data is often the most sensitive, often due to integrity reasons.
- *Business* data could be related to all above. This data relates to the business, e.g., customer data, product data or any other data the company generates.

Data may not only be categorized by its origin, as presented by Shah et al. where the categories of *Human-sourced*, *Machine-generated*, *Open* and *Closed* data are exemplified [37].

Examples of *Human-sourced* data is data generated by a human, e.g., books, spreadsheets, and art, including both analog and digitalized. *Machine-generated* data is on the other hand data which is not generated by a human, instead by a machine. Sensor and log data are good examples, but data as digital art or books can also be this kind of data, as long as generated by a non-human individual.

Further, *Open* data is, as the name suggest, data which is considered to be open, i.e., freely accessible data for anyone to use. This kind of data can come in almost any form, but usually not containing sensitive information, as human or business data. *Closed* data is on the other hand data that is limited in access, for example, limited to the data owner(s) or groups.

Data characteristics can also be essential when assessing the potential value of the data. In a report from the Bennet Institute for Public Policy, with accompanying literature study, value-defining characteristics from both an economic and informational aspect are presented [9]. As for the informational characteristics, which are more of interest in the context of this master thesis, Coyle et al. presented the following:

- What is the data about?
- How general purpose is the data?
- What temporal coverage does the data have?
- What is the quality of the data?
- How sensitive is the data?
- How interoperable and linkable is the data?

Answering these could help to understand and detail the potential value of the data in question. One important note from the economical side is the characteristics of data being non-rival - data does not get used up as more people use it. As mentioned earlier, the non-rival characteristic of data is also an essential characteristic of the value-driving aspect of access to data and sharing of data. Coyle et al. [9], motivate that the parts that can exclude use and access to data can be limited by technology, licenses, governance, and price.

As previously mentioned, Coyle et al. also discuss the openness of data, visualized through the Data Spectrum model detailing the data being either closed-shared or open. In the model, the different data characteristics are mapped with the classification of its closed or open nature. Closed data is the kind of data used for internal access, for example, employment contracts and policies.

2.1.2 Metadata

Metadata is data that is describing other data, usually things like [40]:

- Time/date of creating
- Author
- Description
- Type of data

Everything usually has some kind of metadata, for example, images as they usually store size, color depth, resolution, date and time of creation, or metadata from literature would-be author, title, pages, date, a short description, and a UID. A valuable aspect of metadata is that it can be used for filtering and searching for the data itself. According to Viola and Mookencherry, *metadata* is defined as *"information that describes, explains, locates, and otherwise makes it easier to retrieve and use information resources"* [40].

2.2 Data Ecosystem

As any valuable and tradable good, an interest in publishing, trading, or even selling data is also present. Data users and data providers need a way to interact and collaborate to unlock the potential benefits of sharing data. It is from this notion that the concept of a Data Ecosystem emerges [31].

Data Ecosystem is a relatively unformed concept in academia, and most research can be mapped to the last decade. In 2019, Oliveira et al. [30] conducted a systematic mapping study where 29 studies were selected and further analyzed. The general conclusion from the paper was that Data Ecosystems are gaining recognition and importance but is still lacking in terms of research toward accepted theories, models, and engineering methods for rules, procedures, protocols, and processes to develop, manage and evolve Data Ecosystems. However, the study provides a snapshot of the current landscape where common concepts, essential features, and behaviors of Data Ecosystems are discussed.

2.2.1 Definition and Characteristics

As for a defining Data Ecosystem, based on their findings, Oliveira et al. [30] present a definition of it as being a *socio-technical complex network in which actors interact and collaborate with each other to find, archive, publish, consume, or reuse data as well as to foster innovation, create value, and support new businesses.*

In more recent research, Runeson et al. [35] explored Open Data Ecosystems, and presented a definition of the underlying Data Ecosystem as being:

– *a networked community of actors (organizations and individuals), which base their relations to each other on a common interest.*

– *supported by an underpinning technological platform.*

– *that enables actors to process data (e.g., find, archive, publish, consume, or reuse) as well as to foster innovation, create value, or support new businesses.*

– *Actors collaborate on the data and boundary resources(e.g., software and standards), through the exchange of information, resources, and artifacts.*

By building upon and extending Oliveira's definition, Runeson et al. also take into account an underpinning technological platform inspired by and often present in Software Ecosystems [35, 26]. The similarities and influence between Software and Data Ecosystems are present, sharing characteristics and ideas. Oliveira et al. describe that components frequently found in a Data Ecosystem are *actors, roles, relationships, and resources*, and further detail their interplay in context as: A Data Ecosystem consists of a loose set of interacting actors that directly or indirectly consume, produce, or provide data and other related resources (e.g., software, services, and infrastructure). Each actor performs one or more roles and is connected to other actors through relationships, in such a way that actors, by collaborating and competing with each other, promote Data Ecosystems [31, 30]. In comparison, similar characteristics can be recognized in the software ecosystem where for actors, their relationships as well as sharing and collaborating around common resources are present [26].

Actors are autonomous entities - e.g., an enterprise, institution, or individual. Each actor in the Data Ecosystem can act in one or multiple roles. The roles in a Data Ecosystem are more diversely presented in the literature, although the most frequently occurring role is the *data user*, responsible for directly or indirectly consuming data. Another frequently mentioned role is the *data provider*, responsible for data supply or provision. Related to this, the role of *data producer* exists. In difference to the *data provider*, the *producer* is responsible for capturing and generating data and other responsibilities such as compiling, aggregating, and packaging data. Other not as frequent, but relevant roles presented in the literature are *data re-user*, adding value to the data to be reused, *service provider* and *policies, laws and rules parties*. To note is also the unique role of a *keystone-actor*—an actor who is responsible for driving forces behind the ecosystem as well as providing stability in the unstable environment [30].

An important finding in the study done by Oliveira et al. is the contextualization of Data Ecosystems into four main organizational structures [30]. These are as follows:

- *Keystone-centric* - Actors within the ecosystem gather around a keystone-actor who provides the shared data and orchestrates the ecosystem.
- *Intermediary-based* - A central actor is limited to intermediating data between data providers and data users.
- *Platform-centric* - The actors interface with each other mainly through a platform, e.g., data catalogs and portals.

- *Marketplace-based* - The actor interaction takes place in a marketplace with rules and technical infrastructure that underpins the ecosystem.

Further, Oliveira et al. [30] also classify Data Ecosystems into categorized domains, i.e., the setting or environment in which the ecosystem has emerged. The domains can be categorized into three; scientific, government, and industry. In a scientific Data Ecosystem, the actors are concerned with sharing scientific data in academic or scientific communities, and where one keystone actor is managing the ecosystem. The second domain, government, is connected to Open Government Data initiatives and focus on engaging various actors to promote the usage and publication of open government data. Usually, data providers are public agencies and government units, while data users are citizens. Finally, the last category encapsulates the landscape of industry Data Ecosystems. The authors note that this domain is more blurry to define but states that there is a greater pursuit of innovation, new businesses, and value creation [30].

2.2.2 Variations and Conceptual Model

Although Data Ecosystems are a relatively new concept in literature, two variants, *Open Data Ecosystems* (ODE) and *Open Government Data Ecosystems* (OGDE) who share many similarities, have been defined and further discussed [30].

The underlying factors for the emergence of OGDE are political initiatives and programs for sharing government-related data to the public [30]. Open Data movement and Open Government Data programs, promoting Open Innovation, call for free use, reuse, and data redistribution. Governments and politicians all over the world have started to support the disclosure of Open Government data to accomplish a wide variety of public values [30, 45]. As a response, Open Government Data Ecosystem, a variant of Data Ecosystem encompassed for Open Government data, has attracted research.

The other variant, frequently depict in the literature is the Open Data Ecosystems (ODE), which similarly as to OGDE originates from the Open Data movement and promotes Open Innovation [35]. Differencing from OGDE, ODE are not limited nor focused on governmental data which by regulatory measures often entails open access to the data. Instead, research detailing ODEs are in an industry context, between organizations, although focused on the caveats of collaboration around Open data.

Not until recently, conceptual models for describing a Data Ecosystem and its essential elements have not been present in literature, also requested as further work by Oliveira et al. [30]. Returning to Runeson et al. empirical investigation into ODE as an industry collaboration concept - their findings provide a conceptualization of an ODE.

Stemming from the synthesis of their result, Runeson et al. present their conceptual model into [35]:

- Value - The value of data and value of collaboration around the data. One or the other may be the primary value, but they are highly intertwined.
- Intrinsics - By definition, everything in an ODE relates to the data. However, some aspects are more related to the data's intrinsic or internal characteristics. Among those,

they find the data type and data quality. They also find legal aspects be tightly connected to data, although they also connect to governance of the ODE.

- Governance - The relationship and competition are highly related, as they refer to different kinds of relations between actors. Further, data acquisition also depends on relations between actors in the ecosystem and how they are governed.
- Evolution - The matters of maturity are about further evolution since ODEs are in their infancy and need further research and development.

2.3 Data Catalog

As Data Ecosystems are still emerging, a coherent explanation of what is the specific meaning behind a technological platform is relatively open. However, as discovered from primarily the research from existing platforms in OGDE, one mentioned and in use to support the concept of Data Ecosystems is called *Data Catalogs* [30].

For this thesis work, *the underpinning technological platform* [35] which we intend to investigate is scoped to be one of Data Catalogs.

The term *"Data Catalog"* has only been addressed in the scientific literature for a short while and has recently become more popular. The term is loosely defined as; *A data catalog maintains an inventory of data assets through the discovery, description, and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards, and other data consumers to find and understand a relevant dataset to extract business value.* [43]. As per the definition, the term addresses multiple groups, from data analysts to data consumers, making the term *"data catalogs"* broad. A study done by Labadie et al. found that *"data catalog"* contains multiple functionalities as follows [25]:

- Data discovery - allows users to find data
- Data inventory - data documentation required to present users with a single view on enterprise data
- Administration - user management and system configuration
- Data assessment - metrics such as data usage and data quality
- Data governance - assigning roles and setting up workflows
- Data collaboration - communication and enrich the data within the tool
- Data visualization - displaying data and lineage in a meaningful way and get a better grasp of the data
- Data analytics - insight on datasets and generate new analysis data
- Automation and Machine learning - automated tagging and ingestion

From this research, they found that overall *"data search, data tagging, and workflows"* are the most cited functionalities related to data catalog [25].

Existing Data Catalog Platforms

Many companies have tried to solve the data catalog problem, usually with the starting point of solving data discovery. Most of the existing platforms are open source and free. A few examples of existing platforms are; Datahub, Amundsen, and Secoda. These platforms are all different, however aiming to address the features of *"data catalogs"*.

Datahub originally developed by LinkedIn, describe themselves as a *metadata platform for the modern data stack*, with the key features of data discovery, data observability, and federated governance. The platform uses *"automated metadata ingestion"* or as they also call them *"recipe"* to integrate data sources with the help of their CLI. To be able to integrate a data source, the source must already be supported by the platform, and then the user can write the *"recipe"* which only requires the user to write a few lines of YAML code. These recipes enable users to specify the source and transform the received data with automatic tagging and pointing it to the platform Datahub as the destination. Datahub has support for many of the most popular data sources. Since Datahub is open source, the user could with the help of their documentation, add support for the type of data source they have [1].

Amundsen originated from Lyft, describing themselves as a *data discovery and metadata engine*. Amundsen is open source and wants to help data analysts, data scientists, and engineers interact with data and improve productivity. They do this by indexing data resources to enable page-rank style search. To be able to ingest data into Amundsen, the user has to edit or write a Python script that uses the modules given by Amundsen, and some default configurations for data source integrations [4].

Secoda is a commercial platform with more user-friendly features. Secodas focus is to enable the user to get a complete picture of their data and its context with the main features of searching, organizing, and collaborating on data knowledge. The platform can be integrated with 20+ data sources without having to write any code, sources such as; MySQL, Redash, Big Query, and AWS S3 [36].

Chapter 3

Methodology

This thesis work follows the design science approach, which consist of problem conceptualization, solution design, and validation. The chapter begins to present the 3.1 Research approach, the methodology and motivation to why this thesis followed the design science paradigm. Each phase of the research design paradigm is then further be detailed in the sections of 3.2 Problem conceptualization, 3.3 Solution design, and 3.4 Validation.

3.1 Research Approach

The initial area of interest, and the problem instance for this master thesis, was presented by our supervisor at Axis. Choosing a suitable methodology to conduct our research was of great concern to best perform the research as the overall objective of this research was to, in relation to an increased internal interest at the case company, investigate how to better discover, share and collaborate around data at the case company, also investigate which requirements towards a technical platform is of importance.

According to Aken [3] scientific research can be categorized into the three different categories; *Formal science*, *Explanatory science* and *Design science*.

For *design sciences*, which are often used in engineering and medical sciences, the aim is to understand and improve human-made design in a practical area. The design science paradigm provides a way to conduct research where research problems are formulated and assessed by studying specific problem instances in practice, where research activities are problem conceptualization, solution design, and validation [3].

As for adaptation, Runeson et al. [33] demonstrate how the design science paradigm fits as a frame for empirical software engineering research. Empirical software engineering research

aims to develop and validate practical useful methods, technologies, and tools to help the industry improve software engineering practices. The software itself, e.g., the tools designed to support the engineers and the organizations developing it, are human-made constructs which argues for a feasible adoption of the design science paradigm.

This thesis follow the design science approach in software engineering, and more specifically the one presented by Runeson et al. [33], as visualized in the figure 3.1. Their model of design science spans over the two dimensions of *problem-solution* and *theory-practice* where problem-solving, the practical contribution, takes place in the bottom two boxes, and the theoretical contribution in the two top quadrants. Arrows in 3.1 describe the processes of generating knowledge, which both researchers and practitioners can perform. Further, the activities are performed iteratively across the different domains.

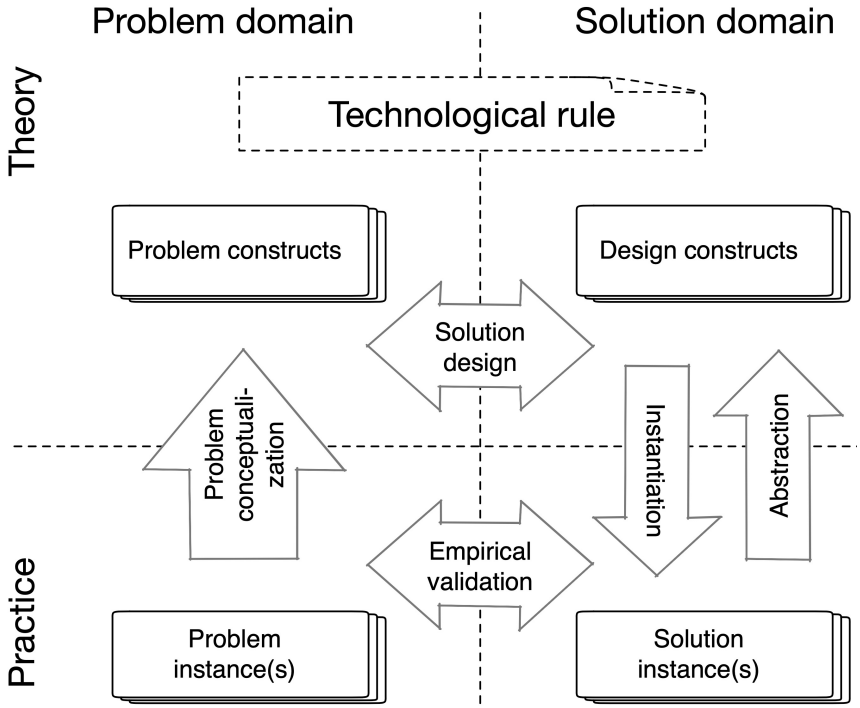


Figure 3.1: Design science model

Activities throughout the model is detailed as:

- *Problem conceptualization* is the process of describing the problem.
- *Solution design* is the process of mapping a problem to a general solution.
- *Instantiation* is the process of implementing the solution in the context of the problem.
- *Abstraction* is the process of describing the key design decision of the solution.
- *Validation* is the process of evaluating the implementation on its problem context.

Discussed in their paper, Runeson et al. [33] note that research contributions under the design science paradigm does not necessarily have to include all of the activities of the design science paradigm. For example, one study may focus on the conceptualization phase while another may cover the full chain from problem conceptualization to solution design artifacts.

Research that focuses on one aspect may build on other research under the design science paradigm.

As for this thesis work, investigating a relative immature research field of Data Ecosystems further the theoretical adoption onto the case company, the research is highly of an exploratory and qualitative nature and will primarily iterate over the theory domain of the problem-solution dimension. The effect will be that instantiation of the solution instances, in practice, will not be done. Instead, the solution instances will refer to a higher abstraction level and consist of the synthesis of requirements derived from the study and the theoretical knowledge in the constructs they build on.

This thesis work also constitutes only one design cycle, i.e., we only iterate the whole paradigm once. Another iteration would of course strengthen the research, and possibly enable us to lower the abstraction level. However, due to the limited time frame of our work, this was not possible.

Besides the relation to context for each phase, the design science paradigm does not prescribe specific methods to be conducted in a research study. However, each phase is more or less suitable for methodologies and corresponding data collected. A presentation of each activity performed throughout our thesis work, with the corresponding area of methodology, can be seen in table 3.1.

To initiate our research, the primary activity was the one of problem conceptualization. For conceptualization a literature and case study was performed, detailed as 1.1 and 1.2. The second phase was the solution design, which in agreement with the design science paradigm was done with high connection to problem conceptualization and corresponds to the activity of synthesis of requirements, see 2.1. Finally, for empirical validation of the theoretical knowledge constructs and returning to the practical sphere, the phase of empirical validation was done by two activities; First, a survey for feedback on the requirements was sent out to the participants of the case study, further an assessment analysis of the requirements against existing data catalog platforms, see 3.1 and 3.2 in table 3.1.

Part	Action	Design science paradigm
1.1	Literature Study	Problem conceptualization
1.2	Case Study	Problem conceptualization
2.1	Synthesis of requirements	Solution design, Instantiation
3.1	Survey of requirements	Empirical validation
3.2	Platform assessment	Empirical validation

Table 3.1: Our research design parts and its counterparts in the Design Science paradigm.

An overview of each activity in relation to its affected research questions, from section 1.2, can be seen in figure 3.2.

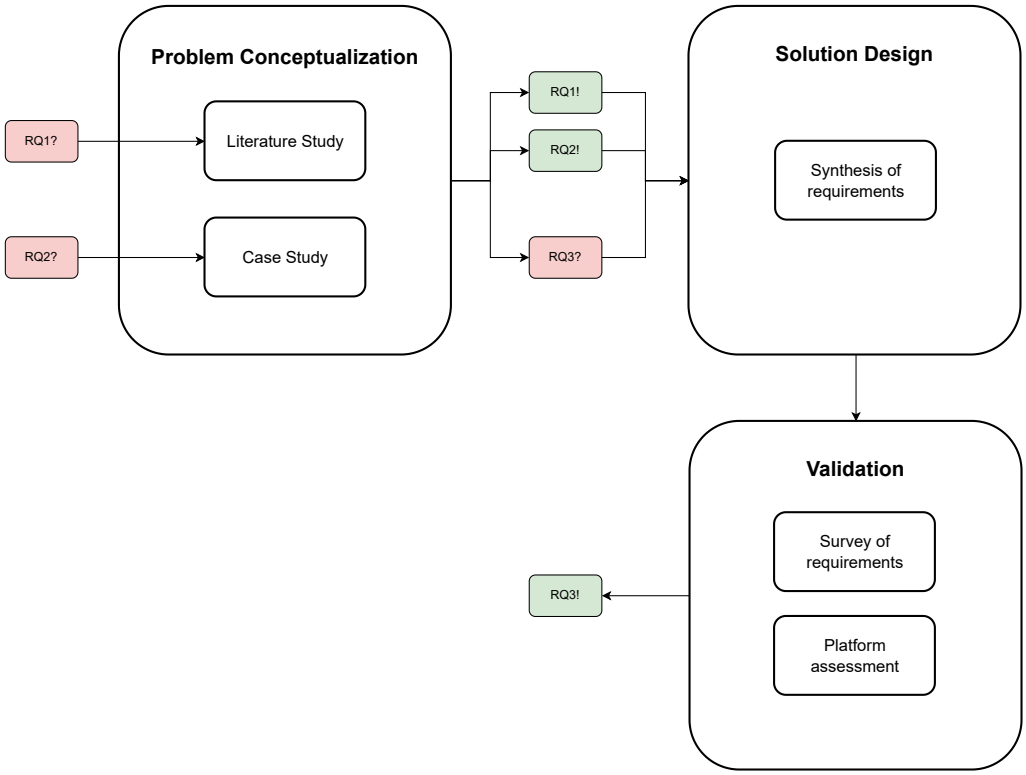


Figure 3.2: Overview of the methodology and how the RQs relates to it.

3.2 Problem Conceptualization

To build knowledge around the problem area and support the later design solution, the first phase and initiation of our research within the realm of the design science paradigm was the problem conceptualization. Following the design science principles, the conceptualization should consist of both empirical and theoretical support with the objective to explore and deepen the understanding of the problem [33].

In fields where theoretical foundation is less mature, as in software engineering, researchers and practitioners may work together to advance and extend the scope of the theory [33]. Similarly, as the research in this thesis was performed in an industry-academia collaboration regarding a relatively immature field of research, the problem conceptualization was done in collaboration with the case company.

Our theoretical support, and the first phase of the problem conceptualization, was gathered by performing a literature study. The aim was to gain a better understanding and document findings related to the investigation of RQ1, also providing theoretical support for the solution design and investigate the problem instance in theoretical context.

The other activity performed for problem conceptualization was to investigate the problem instance further in its practical context. For our research, this meant the exploration of the problem through semi-structured interviews at the case company, allowing us to collect knowledge and insights about thoughts surrounding the theoretical foundation and future

landscape of data sharing in a data ecosystem, providing empirical data for our later solution design.

For guidance and scope, both of these activities were done in high relation to the conceptual model of an ODE, recently published by Runeson et al. [35] and presented in 2.2. The conceptual model helped us to focus our research on the aspects of *Value, Intrinsics, Governance and Evolution* as elements of a Data Ecosystem, further deducting our findings in relation to these.

The conceptualization phase ended with an analysis of the gathered results, respectively.

3.2.1 Literature Study

A literature study is often done to gain a deeper understanding of a subject. It is essential to do the literature study well since this will not only lead to gaining a deeper understating of the subject but also lower the risk of overlooking research on the subject [19]. Moreover, in order to accomplish a good literature study, knowledge about the underlying research around the topic is important. Initiating the literature study, we followed the steps of *search wide, selection, search specific* presented by Höst et al. [19, p. 67].

Search wide was done by exploring the background literature around the area, provided by our supervisors and findings leading up to initiation of the project. This step also involved exploring citations and further searching relevant keywords of surrounding areas. We then *selected* papers, and areas, that we thought were most relevant. Having a better and deeper understanding of the background research revolving the subject, we proceeded with initiating a keyword-search of the research area in order to make our study more *specific*.

Planning for our keyword-search, we referred to the following steps detailed by Thiel [38], suggesting a keyword search should be conducted by:

- Execute keyword-search.
- Select relevant papers from the keyword-search.
- Review the papers abstract for relevance.
- Review the papers for relevance.
- Critical analysis of the results to see if they apply to the new research project.

Before executing the keyword-search a decision to map the keywords to the conceptual model of ODE by Runeson et al. [35], was done. The reason to why was two-fold; First, categorizing our keywords to an area would provide a more coherent approach between us authors, and a conceptual model over the problem domain. Secondly, when further continuing our research endeavour and documenting our findings, the same categorization was planned to be used providing easier mapping within the design science paradigm. Areas from the conceptual model and mapping of keywords can be found in Appendix A.

Once the keywords were established, the actual search was performed in Google Scholar[1] by reading the title of the paper; if relevant, we added it to a list. The search was done individu-

[1] https://scholar.google.com

ally by both authors, adding the results to a shared list. When performing the keyword-search, we added some constraints on the search:

- Result sorted by relevance (matching cases, citations).
- The initial database-search limit to 1 hour per keyword.
- All papers should be accessible through *LU Full-Text Finder*.
- All papers should be Peer-reviewed.

By doing this search, and after merging individual results to a set removing duplicates, we gathered 93 relevant papers. The paper should mainly be relevant for our problem area and the category we had searched for. The next step was to read the abstract of the papers; once again, this was done individually. After abstract filtration, the result concluded that 37 out of the 93 were kept. Here, we started with a more critical analysis. Reading through the papers, discussing interesting findings, and extracting quotes regarding RQ1 and mapping to the conceptual model of the problem was done. The resulting list, in combination with papers from our background research was one of 14 papers; [7, 11, 17, 21, 22, 23, 25, 27, 29, 32, 35, 30, 42, 44], all further presented in the result section 4.1.1 for the literature study.

3.2.2 Case Study

The next phase of our problem conceptualization was to perform interviews at the case company. As detailed by Garousi et al. [16] surveying 101-industry-academia collaboration projects, 75 of these were characterized as case studies. In their paper, they further detailed that industrial case studies usually apply either the exploratory or improving type, or both. For empirical support in conceptualization of the problem, we set out to perform an exploratory case study. The reason to why a case study was most suitable for this thesis was two-fold; First, a case study is a good way to investigate and examine a particular case within real world context, and as we wanted to explore the problem instance deeper and from different aspects at the case company it was well suited. Secondly, case studies also support a rich plethora of options to conduct the study [16]. For example, it can involve interviews, focus groups or observational studies.

For us, the case study was done by conducting interviews with a selection of participants at the case company. The motivation behind interviews being best suited was related to our exploration of RQ2, deepening our understanding of the problem area at the case company and providing empirical support for the later solution design. As RQ2 revolved around exploring the concept of Data Ecosystem in an intra-organizational context, interviews allowed us to best approach the potential actors connected to the problem area, i.e., stakeholders of problem instance, working with data at different departments of the organsiation.

Case Description

As mentioned in section 1.1, Axis is a large company with more than 3800 employees where the most significant part is located at their headquarters in Lund; this also means that Axis has a lot of different departments.

Axis sells network-connected products primarily through partners. To improve their products and make their customers happy, they gather data through, e.g., customer analytics opt-in, testing on-site, feedback, and more. These collection points generate a lot of incoming data located throughout Axis. Many of the teams gather some form of data, not necessarily connected to the products themselves instead it can also be around processes and tools, which leads to the data varying in types and formats

The departments at Axis are often self-driven with limited synchronization and centralized control, which has its benefits and challenges. However today, when the value of data has been observed, an interest in wanting to work more together around the existing data has started to appear.

Figure 3.3 presents an overview of how the department DDM collaborates with data as of today, from DDM's perspective. Data is primarily gathered from products at opt-in customer sites, further to be analyzed and provided to stakeholders across the whole organization. The figure 3.3 also detail the incoming data generated on-site, primarily generated inside and used by Research and Development (R&D) departments.

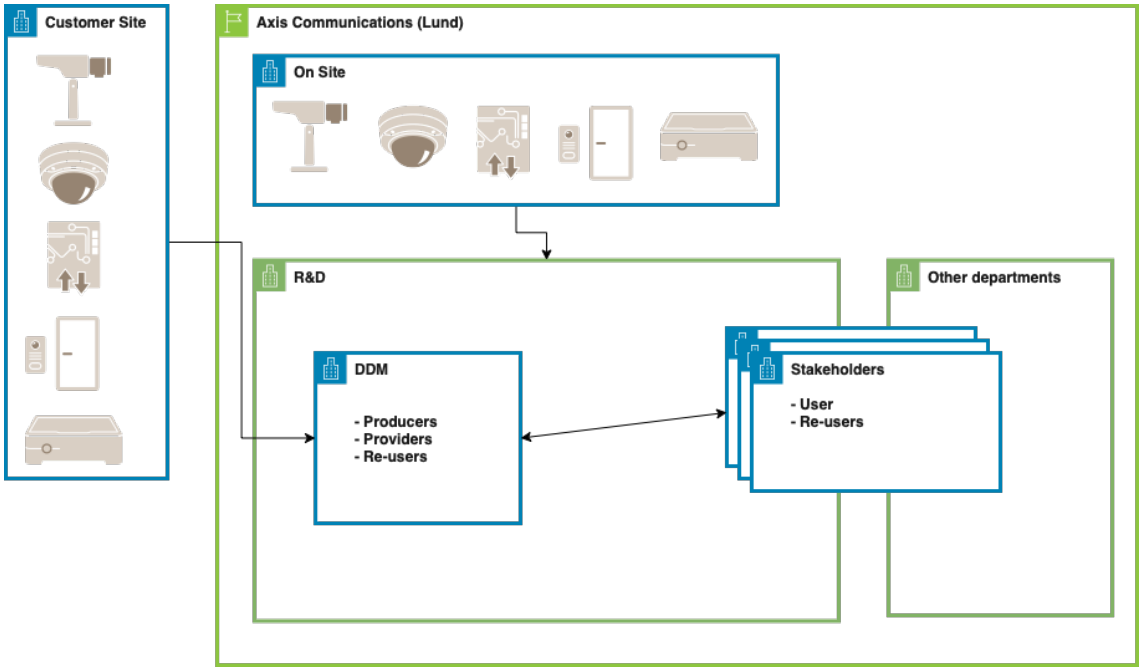


Figure 3.3: Overview of the current data input and collaboration within the organization, from DDM's perspective.

Interviews

The interviews were planned to consist of the activities *Research and design of interview material*, *semi-constructed interviews* and *analysis of results*. Given the collection of qualitative data, we planned to follow the qualitative analysis, which consisted of four consecutive steps; data collection, coding, grouping, and conclusions as suggested by Höst et al. [19, p. 114].

Data collection is the process of collecting data through interviews, observations, transcriptions, etc. Differing from, and a mixture between closed and open interviews, semi-constructed

interviews allow the interviewer to follow a pre-defined framework of themes and questions but still be open to discovering new areas [39]. The structure encourages having open-ended questions and room to probe and follow up on the answers to discuss and further explore the participant's proposition.

As the objective of the interviews was to investigate and explore the problem area, without limiting the exploration of surrounding areas, we decided that semi-constructed interviews were well suited.

Before conducting the interviews, a process of preparation had to be done. Following the guidelines by Turner III [39] the process consisted of the selection of participants, pilot testing, and constructing effective questions. As for selecting the participants, we conducted a criterion-based sampling. Here we created criteria based on the ecosystem roles found in the literature. In collaboration with our supervisor at Axis we established a list of potential participants matching the criterion. Each participant were approached by email, providing a description of the interview and setting, question of participation, a suggested time slot and, finally, background information about our research and that of Data Ecosystems. This was done in order to address the purpose of the interview, receive consent for participation and introduce them to relevant and necessary background theory around the area of research. The resulting list of participants can be seen in table 3.2.

ID	Role	Ecosystem Role	Department
P1	Data Manager	Provider	1
P2	Data Scientist	Re-user	1
P3	Data Engineer	Producer	1
P4	Data Engineer	Provider	1
P5	Data Engineer	Producer	1
P6	Legal Advisor	Law & rules party	2
P7	Data Manager	Keystone	3
P8	Data Manager	Keystone	4
P9	HW Engineer	User	5

Table 3.2: A presentation of actual role, ecosystem role and department ID of the participants

Originating from the theoretical support established, and following the same conceptual model, we chose to create a more extensive set of questions related to each area of interest. The procedure to produce practical questions was done by iterating through a set of questions gathered from the literature, grouping, and refining down to a smaller, final set. Two to three questions per area of interest were selected, leaving time and room for probing and follow-up questions. Slight alterations of the questioning were done, depending on where the participants currently worked in the data landscape. This was done with care and changed primarily on how we were wording the questions. The final set of question, with marked alterations, can be found in Appendix B.

In this thesis we conducted one-on-one interviews with each participant, in a mixture of face-to-face and via video calls due to the pandemic. The setting of one-to-one interviews was detailed in the initial e-mail, where each person was given the possibility to accept or

decline and further state if face-to-face or via video was preferred. All sessions were held in an isolated conference room in order for the interviews to be done in a setting with little distraction. As for the video calls we, as interviewees, located ourselves in a conference room while the participant called in.

The implementation of the interviews was performed in the following way:

1. The participant was welcomed and presented with the purpose of the interview.
2. The participant was asked if they had any questions regarding the background material or interview.
3. The participant was presented with the terms of confidentiality, i.e., recording and further analysis and usage of anonymous material.
4. The interview was performed.
5. Recording was checked and uploaded to safe storage for later transcriptions.

Before executing an interview, we divided the work into one responsible for taking notes while the other focused on leading the interview. Both participated in the further exploration of interesting answers.

Following the interview, transcription of each interview were conducted. After the transcriptions were done, we continued with the coding and grouping mentioned previously in section 3.2.2, resulting in the following tables which can be seen in Appendix C.

Coding

Coding is the step where important quotes and information are highlighted and connected with one or more keywords [19, p. 115]. This was done in a combination of a deductive and inductive approach [13]. Initially, we performed a top-down deductive coding approach where both of us went through the transcriptions together and highlighting important quotes in relation to the sub-subjects provided by the conceptual model by Runeson et al. [35], resulting in that large parts of the interview with unrelated information sorted out.

Following the initial deductive coding based on the conceptual model, we proceeded with an inductive coding where the quotes were analyzed and connected with one or more keywords which we discovered in the data. These codes are detailed as the smallest components in the results presented in Appendix C.

Grouping

Grouping takes the keywords of the coded quotes and groups them into more significant categories [19, p. 115]. This was done by taking all the unique keywords per sub-subject from the conceptual model and grouping them into different categories. The categories were chosen to relate to the initial deductive coding, by subject and sub-subject from the conceptual model. The final groupings are the internal titles to the relating sub-subject in Appendix, e.g., *Knowledge* being a group related to the sub-subject *Value of data* seen in Appendix C.

Conclusions

The conclusion is based upon the groupings where we can draw conclusions from it [19, p. 116]. Here we could see what challenges the case company has and platform-specific requirements that were more predominant. These will first be presented and further detailed in the result section 4.1.2 for the case study.

3.3 Solution Design

The solution design phase of our work consist of the one activity; *Synthesis of requirements*. For the synthesis, the plan was to derive problem-solution pairs between the two respective parts described in section 3.2 with the goal to investigate and further answer RQ3, i.e., *what requirements should a technical platform fulfil to address the challenges found in RQ1 and RQ2?*

The resulting artifact from this phase was ambitioned to be a set of derived requirements toward a technical platform, for supporting and mitigating the challenges found in the theoretical and empirical knowledge constructs built through problem conceptualization.

To quote Cruzes et al., *"A research synthesis is a collective term for a family of methods to summarize, integrate, combine, and compare the findings of different studies on a specific topic or research question.* [10]. In the context of this research, the synthesis would be to integrate and combine our two data sources on the research question of RQ3.

Similarly, as for the case and literature study, both being qualitative, choosing a method suitable for the synthesis of qualitative data was motivated. Detailed in Cruzes et al. [10], the three most frequently used methods used for the synthesis of case studies rich in qualitative data in software engineering research are; narrative, thematic and cross-case synthesis. The article also compares the three methods, detailing their characteristics, further effect on the outcome.

Given that the data material for synthesis was limited to one case study combined with the literature study, the analysis method of choice was to use one of narrative synthesis.

A narrative synthesis is generally seen as the most flexible method, where primarily words and text are used to condense and explain the findings, and where the outcome is progressive linking to form a chain of reasoning [10]. Also discussed by Cruzes et al., a drawback of the narrative synthesis is the general lack of transparency given a wide range of specific methods for synthesis that can be used and lack of insights in the procedure, especially in relation to other synthesis methods as thematic or cross-case. Further discussed is the drawback of lacking a sampling criterion for the data of synthesis, making the collection highly dependent on the convenience of the analyst [10].

Despite the drawbacks, a narrative synthesis was deemed suitable for conducting the solution design phase with the motivation of the data used for synthesis previously collected within the same study, hence the chain of evidence from the final synthesis to the original data could be depicted through the problem conceptualization, as an important for any qualitative analysis [34]. Further, to combat the lack of transparency of methods performed, the synthesis was planned to be done in an iterative manner where each data source would be preliminarily

synthesized by themselves, followed by exploring the relationships within and between the studies.

Another aim of performing the synthesis through three separate iterations was to better decouple the material from each other, giving separate and transparent outcomes from each. Also, the separate iterations were planned to mitigate other factors affecting the outcome as potential variation and quality in the different materials and the impact of the goal of synthesis, as detailed by Cruzes et al. [10]. This is because it would not be until the third and final iteration the mapping between the data sources, searching for problem-solution pairs toward our goal of the synthesis, would be performed. All three iterations of the synthesis were done by both authors individually where grouping and coding of the findings from each were documented to finally; be explored together as a reiteration of the third.

For the **first iteration**, a re-analysis of the axial coding results from the singular case study was done as a selective coding of the challenges and benefits [20]. As for the benefits, these also included recommendations of already existing practices and ideas for the future. Here, as previously done for both the case study and literature study, the findings were mapped to the related element of the conceptual model. Resulting from the first iteration of synthesis, a thematic set of challenges and benefits, including recommendations, had been done.

For the **second iteration** of synthesis, similarly as for the case study, the challenges and benefits, where the latter included recommendations, from the literature were sought after through selective coding. Once again, the analysis followed the elements from the conceptual model, where findings were mapped to the area in question. The iteration resulted in a set of challenges and benefits, also recommendations, to be noted.

In the **third iteration**, the two sets of challenges and benefits were analyzed together, from which a final abstraction and formalization of requirements were performed. Discovered problems as challenges, and benefits, raised from the synthesis of the case study were often mapped to a solution in the form of similar challenges or recommendations, noted as benefits, from the literature. The contrast between the two was also noted and influential towards the abstraction of a requirement. Once done, a final set of 14 requirements towards a technical platform had been formalized.

Finally, for **reporting** the synthesis, the narrative synthesis of *"tell a story"* for each finding leading to the formalization of each requirement was performed. Following the same order as the iterations, the findings from the case company were first presented following the findings from the literature. Finalizing each narrative presentation, the mapping to the final set of the requirements addressing the findings were presented.

3.4 Validation

According to the design science paradigm, the final phase of research under the design science paradigm is to empirically validate if the proposed solution address and solves the problems identified, i.e., the requirements toward a technical platform mitigates the problem instances found in the conceptualization of the problem. As discussed by Runeson et al. [33], the context should define the scope of validity in design science research, in terms of the abstraction levels corresponding to the ones in conceptualization of the problem.

For this thesis work, as both the problem and solution instances were explored at the relatively high abstraction level of the case and literature study, naturally returning to the same context and validating the solution instances was done. Further, as for the resulting requirements toward a technical platform, we also wanted to validate the feasibility of use in practice by evaluating the requirements against existing data catalog platforms.

Activities to perform as validation were therefore two-fold: First, surveying the stakeholders of the problem instances at the case company was done to evaluate the feasibility of the requirements as a solution in the case study context. Secondly, a validation and instantiation of the requirements onto a set of existing technical platforms, more specifically scoped to data catalogs as described in 2.3, were performed.

The following sections first detail the methodology used for surveying the requirements against the stakeholders at the case company, followed by presenting the method for evaluating the requirements against existing data catalog platforms.

3.4.1 Survey of Requirements

Due to the problem instance originally being initiated and further investigated through a case study at the case company, evaluating the requirements in the same context was motivated. The method for performing this was to survey the stakeholders at the different departments by conducting a questionnaire. Motivation to use a questionnaire was primarily to collect the data of opinions and be time efficient for all parties involved.

Presented by Kitchenham and Pfleeger [24] the initial phase of designing personal opinion surveys is to establish the objective with the survey. The objective of our survey was to evaluate if the proposed requirements toward a technical platform were feasible as a solution to the problem instances discovered, in the opinion of the stakeholders. Further, we also wanted to investigate the perceived relevance of each, potential misconceptions or further elaborations regarding the specific requirements found.

For design, the survey was a self-administered questionnaire as detailed by Kitchenham and Pfleeger, i.e., the recipients are responsible for reporting and submitting the data without interaction by us. The questionnaire was digital, making it easier to distribute, and the tool of choice was Microsoft Forms[2]. Due to it being self-administrated, the first design element to be included was a description of the objective and layout of the form.

[2]https://www.microsoft.com/en-us/microsoft-365/online-surveys-polls-quizzes

The next step was to design the actual questions. Each question would correspond to a requirement alongside an exemplification of each. As the objective of the survey was to capture the opinions of the stakeholder regarding the proposed solution instances, we decided to use an ordinal scale for the collection of answers. Generally, it is better to use an ordinal scale for attitudes and preferences, in contrast to, e.g., yes/no question or response categories [24]. The ordinal scale used was an agreement scale, in which the scale was from strongly disagree to strongly agree, further balanced by five points with a neutral standpoint in the middle.

Another design decision of the questionnaire was to add two open questions allowing free text. The reason for this addition was to capture any further elaborations regarding one or more requirements, potential misconceptions or other thoughts around the survey or result.

Before the survey was distributed, a pilot test was performed to capture design mistakes and provide feedback. An addition resulting from the pilot test was to add an exemplification of each requirement for an easier understanding of its meaning. The final questionnaire can be seen in Appendix D.

Once ready to be distributed, the target population was set to be all the participants who had previously been interviewed in the case study; hence a list of participation was the same as presented in the case study, see table 3.2. The reason why the case study participants were approached for evaluation was primarily due to the objective of the survey. As the stakeholders of the actual problem instances, evaluating the proposed solution's feasibility was figured to be best performed by the same set of participants. Further being a relative juvenile area of research, the participants had already interacted with the background theory and discussed the subject.

The survey was distributed by email, including a friendly introduction, presentation of the objective, request for consent and, finally, a link to the digital questionnaire. The questionnaire was planned to be open for five consecutive days, or as all participants had answered the survey.

Finally, for analysis of the requirements, the results were two-fold; For each requirement, the quantitative data from the survey was presented. Secondly, findings from the final two open questions were detailed and discussed.

3.4.2 Platform Assessment

As for the platform assessment, the main objective was to assess the feasibility of use of the requirements toward a set of technical platforms. Detailed in section 2.3, from our background research we had found and scoped our existing platforms to the three *data catalog* platforms *DataHub*, *Amundsen* and *Secoda*. These were selected based on findings from the background research around Data Catalogs, and further motivated by our supervisor at the case company to be suitable candidates for future establishment.

The aim of the validation was to see if the requirements would be feasible for use in the practical context, and the scope was limited to only using the given information in the form of documentation, a preview of features and in some cases, demos on their website to validate if they had the given requirement. We did not set up any of the platforms to validate them locally. However, for evaluating the feasibility of usage and further implementation of the re-

quirements in its practical context, we still wanted to perform this analytical activity against the selected platforms.

The steps of validation were as follows:

1. Individually go through the documentation for each platform, assessing if it fulfilled each of the requirements.
2. Discuss our findings and note the final decision.

Both of the open-source platforms, *Datahub* and *Amundsen*, had developer documentation of the platforms with how they worked, since being Open Source, this documentation could be out-of-date with the current version. For *Datahub*, they had developer documentation and a demo on their website, making it easier to see what they were offering and how they tried to solve different problems. Meanwhile, *Amundsen* only had the developer documentation, which could be a bit hard to navigate. *Secoda*, a commercial platform, had no developer documentation on what it had and how it worked. They instead used previews of the platform's features, making it harder to confirm what their platform had and did not have. Nevertheless, being commercial, the previews of *Secoda* can be argued to put more importance on being up-to-date and more user-friendly displayed for their customers to see what they were buying.

Chapter 4

Results

4.1 Problem Conceptualization

As for problem conceptualization, the following section presents our results from analysis of the literature study, followed by the case study.

4.1.1 Literature Study

The results gathered from the literature study will follow the same categorical mapping to the conceptualization presented by Runeson et al.; Value, Intrinsics, Governance and Evolution [35].

Overall, the papers investigated include the initial findings from the background papers and results from the keyword search. Papers gathered were primarily from research domains of Data Ecosystems, Open Data Ecosystems, Big Data, Collaboration around Machine-Learning and Data Management.

Value

Detailed by Runeson et al. [35] their synthesis of the value in an ODE were (1) an ODE is driven by the value of data or the value of collaboration and, (2) the value of collaboration is impacted by the competition between actors. Even though seen as a value in the Data Ecosystem, collaboration around data poses some challenges.

In a recent publication Nahar et al. [29], interviewing 45 participants from 28 organizations investigating key collaboration challenges in ML-enabled systems, they found that the teams responsible for building the model are in most organizations not the team that collects, owns,

and understands the data, making data a key collaboration point in an intra-organizational environment. Conclusions from the paper were that communication, documentation, engineering, and processes all posed as present challenges for collaboration around data.

In 2018, Kim et al. [23] performed a survey on a large number of data scientists to better understand their work and present challenges, tools, and activities used, and found similarly as to Nahar et al., collaborative issues. Kim et al. research suggests that, for effective interdisciplinary collaboration, formal training, standardization, clarifying the goals, and understanding the caveats of data were key aspects to address. Regarding formal training, education was in both papers brought up as a key aspect in combating challenges around miscommunication and achieving effective knowledge transfer [23, 29].

Establishing relationships where expectations are clear and a formal contract, specifying data quantity and quality expectations was detailed to be a beneficial action according to Nahar et al., as well as common documentation understood by all, e.g., data item definitions, semantics, or schemas [29].

The study by Kim et al. [23] addresses another important area in a Data Ecosystem, which is the characteristics of the data itself. To overcome challenges around the actual data, e.g., poor data quality, missing or delayed data, or need to shape the data to fit a diverse set of tools, data scientists suggest consolidating heterogeneous tool suites and creating data standards for instrumentation, moving toward standardization. The authors suggest the need for engineering practices such as data quality checks handling the evolving nature of data, deployment automation, and testing in production were detailed as examples of these.

Intrinsics

As proposed by Runeson et al. [35], the intrinsics of data plays an important role in an Open Data Ecosystem. Conclusions were that the type of data and its characteristics impact the degree of openness, standardized meta-data and domain models are core quality attributes for data, and legal frameworks must be developed to support the evolution of ODEs.

Regarding the types of data, the sensitivity of data seems to impact the willingness of actors to share data. Classification of what is sensitive data may vary, although human data as health data is often detailed to be the most sensitive and its use regulated [35]. Another sensitive data type mentioned by Runeson et al. is the one of business data. Business data may vary depending on the business but usually include data connected to the customers or about the business itself [35].

Immonen et al. [22] investigated certified open data in an ecosystem, they detail that due to private data are often the customer's data, or information collected about customers, the management of privacy become crucial. However, sensitive data is in some sense communicable - i.e., when combining data, the data can become private when private elements are added to the data set. Where this can be imagined to pose a challenge is concerning the notion that processed data may involve less sensitive details but may, on the other hand, also be more valuable, if combined with other data sources, as found by Runeson et al. [35].

To be able to assess the openness of data in the context of the open-closed-shared spectrum, recent research done by Enders et al. [14] present a conceptual framework to guide the

decision-making. The authors suggest questioning the data in relation to the metrics of coreness, currentness, extent, granularity, interoperability, and quality of the data. Further, five decision criteria are suggested depending on the metrics; competitiveness, data misappropriation, innovation opportunity, legal, and privacy. Another factor, noted on a macro-level, is the necessity of data providers' knowledge about the types and formats of data that are demanded by entities in the ecosystem.

In 2016, Custers and Uršič [11] discussed the balance of big data and personal data protection, explaining that the regulations around EU personal protection directive (GDPR) require that data must be collected for specified, explicit, and legitimate purposes and not further processed in a way incompatible with those purposes. Their findings also point out that it is a risk to limit the use of data, as regulatory directives restrict the use.

Concerning privacy, Boeckhout et al. [7] discusses the challenges and opportunities raised by FAIR principles and how to enact them responsibly, suggest that metadata may be a solution to challenges around privacy of data. FAIR stands for Findability, Accessibility, Interoperability, and Reusability and was in 2016 detailed as guiding principles and practices for data stewardship in the life sciences [41]. The principles of FAIR have since set out to provide guidelines to improve the FAIR elements of digital assets, emphasizing machine-actionability [15]. Even though Boeckhout et al. article resides in the context, further influenced by the data in life sciences, working with sensitivity and privacy around data is relevant for outside the life sciences. Debated by the authors, FAIR principles and metadata standards could help facilitate compliance with the principle of data minimization by allowing for an assessment of which data to reuse based on an analysis of metadata, which by and large should be non-personal [7]. However, the GDPR also aims to ensure that any data use is stipulated as clearly as possible in advance, which still shrouds the issue of well-defined use cases and suggests further research.

In 2020 Gelhaar and Otto [17] found, investigating the challenges in the emergence of Data Ecosystems, that standardization initiative and the agreement on standards to be a key factor in the development phase of Data Ecosystems. Especially one crucial aspect of standards in Data Ecosystems is to enable interoperability and enable efficient data sharing. Interoperability is important to ensure non-discriminatory data sharing between the actors and, on the other hand, enabling further data processing through a common understanding. Hence, standardization initiatives and the agreement on standards can be critical factors in the development phase of Data Ecosystems.

With regard to standardization around the data, Yoon [42] investigated failed data reuse experience, found vital issues was the one of access and interoperability, further suggesting that processes should also address interoperability in data formats and software for successful data reuse. Another conclusion by the author was a general lack of support in reusing data was frequently apparent, emphasizing the need to develop a support system for data re-users.

In a more recent paper by Immonen et al. [21], with a focus on quality aspects of the data in an ODE, detail that poor and unknown quality has widely been recognized as one of the significant obstacles for open data utilization. Since open data can be almost anything and originate from different data sources, data quality becomes the key issue. However, quality is challenging because it is difficult to judge without considering the context or situation. Conclusions from the paper is that key activities revolving around securing quality is bring-

ing data from trustworthy sources, transforming it to the acceptable form of the ecosystem, validating it against its intended usage of each service provider, monitoring the data sources and the usage of the data, and continuously evaluating the quantified value of the open data service. The authors also suggest establishing Service License Agreements (SLA) with the producers, as they, e.g., can be transformed into tactical rules used for quality evaluation.

Governance

Research around the areas of governance, management, and coordination of Data Ecosystems is in its current state sparse, as also highlighted by both Runeson et al. and Oliveira et al. [35, 30].

Recently though, Lis and Otto [27] performed a case study around Data Ecosystem governance, explicitly investigating the difference between intra- and inter-organizational data governance. Besides their findings, the authors also explain intra-organizational governance, where the decision rights for the management and use of data manifest within the organizational structures and hierarchies, which may vary between organizations. Originating from IT governance, data governance has remained relatively stable around the areas of provision decision-rights, roles, and accountabilities for the management and use of data [2].

Comparing intra- and inter-organizational governance, Lis and Otto found that for a Data Ecosystem, the importance is for the actors to share and understand their and other roles within Data Ecosystems, as the range of data governance mechanisms may vary depending on the position the organization is operating. They also highlighted the importance of trust and exemplified ensuring data security, data availability, and data integrity as key challenges. Moreover, the trust may be affected by each organization's understanding of the playing field. The organization in control of the platform infrastructure possesses more control mechanisms and can influence user interaction and data management. Finally, an important finding was that if the organization's internal foundation of maturity were good, the success of engaging in inter-organizational data collaboration would increase. Suggestions of key areas to address are establishing knowledge around which data exist and are relevant within the organization; who is responsible or can provide information related to these data assets; how the data is used (both internally and externally), and under which conditions data can be shared with whom and where.

Gelhaar and Otto [17] in their study around challenges in the emergence of the Data Ecosystem, in a similar fashion, detail the necessity and challenge around trust and transparency between the actors. To overcome this challenge, they suggest that the organization should provide a transparent presentation of how and by whom their shared data is used. The transparency of the data usage can include the complete history of the data, from its creation to its transformation and usage.

Turning away from specific ecosystem governance, the need for its presence can still be found in literature around the Data Ecosystem. Zuiderwijk et al. [44], presented in 2014 research around essential elements of ODE. Their findings conclude that an open Data Ecosystem should capture searching, finding, evaluating, and viewing data and obtaining information about the licenses related to the data. Further, the data is also suggested to be obtained by request, e.g., users can request data from data providers and provide them with feedback

after they have used the data. Feedback and discussion of usage are key elements of a Data Ecosystem, which should support and enable interpreting and discussing data and providing feedback to the data provider and other stakeholders. E.g., a person who has used a dataset could leave a message on an open data platform to inform other users of the dataset about certain particularities or post a question about the data. Furthermore, to facilitate maturing of the ecosystem, the need for user guides showing how the data can be used, a quality management system and different types of metadata for being able to connect the elements are suggested as elements to include.

In the context of governance, Boeckhout et al. [7] research on the usage of the FAIR guiding principles for data stewardship also detail some suggestions. Conclusions from the paper are that the FAIR principles, in regard to data governance, enforced three important questions to address. First, the FAIR principles stress the importance of metadata and metadata standards in data stewardship. A key message of the FAIR principles is that metadata and metadata standards should be articulated and made publicly available to the greatest extent possible. Second, many aspects of data stewardship are suggested to be assisted and executed by computers, therefore setting a crucial prerequisite for automation. Thirdly, the FAIR principles call for explicit, well-defined, and readily available terms and conditions under which data are shared or made accessible. As previously mentioned, the FAIR principles are primarily present regarding data within the context of life sciences; hence application may not be directly transferred to other contexts. However, governance of data shares similarities regardless of context. As life sciences revolve around sensitive data, much like other data types such as business and human data, findings might be of interest in the context of Data Ecosystems outside the realm of life sciences.

Evolution

Finally, as Data Ecosystems evolve, the area of evolution encapsulates the essential aspects as the ecosystem grows and matures. Still in their infancy, in regard to ODE, Runeson et al. [35] proposed that further research around the integration of ODEs into an organization's business model as well as tools to support such an ecosystem and enable data sharing should be developed and standardized.

In another study concerning the usage of OSS principles for data sharing, Runeson [32] notes that there is a general lack of maturity within organizations when it comes to sharing data; this as that the companies stored data in different municipalities, use of various technological platforms and APIs. The paper also brings up that because organizations are not used to sharing data with others, processes or procedures for how to act are not well-defined, which poses a challenge. However, as presented by the authors, some evolutionary directions to consider can be found in one of the more established collaborative landscape of OSS: To quote Runeson *"The development of, e.g., Git as a distributed configuration management tool, Jira for issue management, and Slack for communication, have contributed significantly to making OSS widely adopted."* [32]; hence the maturity of technical platform addressing the current challenges and potentials of Data Ecosystems can only be seen as an essential evolutionary step.

Labadie et al. [25], presented in their explorative study assessing data catalog initiatives the finding that after a company has implemented a simple Data Ecosystem, the employees must adopt a *"data mindset"*. With *"data mindset"* they mean that the employees should know about

the data, where and how to find it, and what it means. They suggest that a data catalog following the FAIR principles could help with this mindset, but more implementation research should be done.

4.1.2 Case Study

In this section, our final conclusions from the interviews are presented. For the case study, the same conceptual model was followed for analysis, mapping our findings to the domains of the value of data, intrinsics, governance, and evolution. A visualization of the final mapping of themes can be seen in figure 4.1. Full coding from analysis can be found in Appendix C.

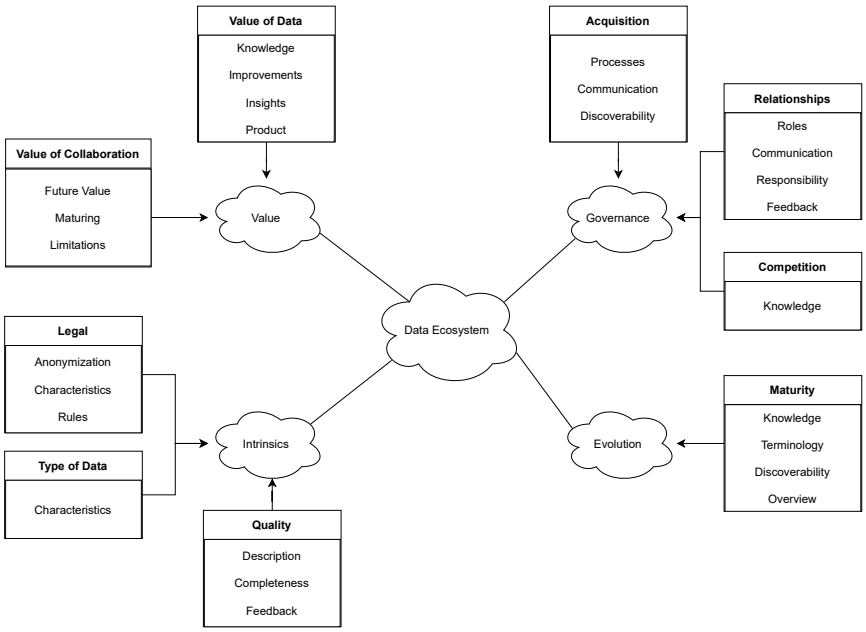


Figure 4.1: Final mapping of keywords from the case study.

Value

For value we had two sub-categories; *Value of Data* and *Value of Collaboration*.

In figure 4.1 we can see that *knowledge*, *improvements*, *insights* and *product* was the concluded themes for *value of data*.

From the interviews, a majority of the participants mentioned that the value of the data was connected to increased knowledge. Knowledge was described in terms of new improvements, insights, understanding, and better overview. More notably, the notion of increased knowledge in some way was mentioned by actors throughout the whole ecosystem. One participant stated that *"For Axis, the value is knowledge and future improvements"* which encapsulated the overarching theme well.

Regarding when the value was generated showed two more distinct perceptions to either be at a later stage or in the current. For the producers, providers, and re-users, the answers presented the idea of value being generated in a later stage. In contrast, the answer indicated a generation of value in the present for the end-user. One of the providers described it as *"In order to be a provider, you need to be a user as well [...] You need to understand if it works, and if the data has a value"* while the end-user described the value of data as *"Understand more about how the product and specific components are doing, better or worse"*. Another confirmation of the symbiosis between the providers and users were the re-user within the team stated that *"Half of the features are made by us [own initiative], and the other half is made by orders from end users"*

A common description of the value of data was coupled between the data collected and the outgoing value. All three keystone actors from the different departments discussed aspects of data; quantity, currentness, and completeness affecting the value. To quote the keystone actor from department 4: *"more assets in the data set will provide a bigger insight into what the next step is and give the customer a better product"*.

Value of collaboration detailed a large quantity of keywords for limitation, from the results in figure 4.1, we can see that *future value*, *maturing* and *limitations* were the final themes.

Regarding the value of collaboration, there were more diverse opinions and ideas of the current and future. Unsurprisingly, the most coherent opinion was about the value of collaboration, primarily the potential to increase the value of the data. A theme of benefits of collaboration was associated with maturing, both as interdepartmental and as an organization. For the actors within the DDM, maturing in transparency, spreading knowledge, and receiving more feedback from their end-users were noted as examples of important factors to increase collaboration. The re-user of the department described it as *"The value is dependent on the end-users, and will probably increase once more users get to know about the data"* pointing out the dependency of collaboration between providing and using data.

The level of maturity was also a theme discovered when looking toward limitations mentioned concerning collaboration. Participants from all departments described a limiting factor today as the feeling of working in silos and non-standardized ways. A keystone actor from outside the Research and Development department stated that *"the bigger Axis get, the more silos exist [...] You own your problems"* and further *"in order to not hinder us we need to work from a common ground in relation to data"*.

Intrinsics

For intrinsic we had three categories; *Type of data*, *Quality* and *Legal*.

From the results in figure 4.1 *characteristics* was the singular theme noted for *type of data*.

Types of data detailed by the actors from the departments were; dept. 1 - telemetry and logs, dept. 2 - did not mention any specific type, dept. 3 - business data of varying types and finally for dept. 4 - image and video, see table C.3. From the interviews, we could see that all the external data gathered are from customers who have accepted them by "opt-in". Only looking at the DDM team, then according to a participant *"today we have 700 data points"* and all of these data-points are different and collected at different rates depending on how often the data change and what is needed; *"every second week, at least changes"*. In contrast to department

3, they collect a lot of completely different types of data from different systems; *"we get data from 20 systems, all with different column"*, this is a significant issue according to a participant: *"our issue is the sheer volume and variety of incoming data"*.

When analyzing the area of *quality*, more variety in the answers were discovered, although, *description*, *completeness* and *feedback* were the concluded themes, see figure 4.1.

Without having some sort of quality requirement of the data, the data can be rendered useless in some instances according to three of the interviewees, a provider, re-user, and a keystone actor; *"sampling frequency once an hour can be too bad according to a stakeholder, which makes the data unusable"*, *"missing values can cause problems, but sometimes it is ok"* and *"right type of FPS, no lost packages, non-corrupt files, and correct resolution [...] Completeness of data as far as possible"*. To be able to give feedback about the quality back to the producers is an essential factor for a user and keystone actor and is sometimes not that easy; *"feedback from the actual user to iterative correct quality issues"* and *"feedback by showing how bad the quality is in order to get a change, we depend on human knowledge"*. In contrast to the user and keystone actor wanting to receive good quality data, the producer does not care how the data looks: *"raw data is allowed to be any format, as long as it is used"*. This compared to the providers saying that they are *"not responsible for checking the quality of data"*. However, they are still doing some quality filtering for the user/re-user to *"remove the outlier above some level"*.

Legal had *anonymization*, *characteristics* and *rules* as the concluded themes, see figure 4.1.

Regarding the interview topics of legality concerning the data, we could see that there was no real responsibility on anyone as long as the data stayed anonymous and within Axis. This was also confirmed by the legal advisor: *"anonymous data, within the company, does not have any legal stipulations"*. Even though there is generally no responsibility on anyone, DDM uses contracts that the users have to sign, *"a contract where the responsibility is on the end-user"* and *"stakeholders are responsible for complying with the use-cases"*. This contract mainly exists so that the user is responsible for not connecting the data to sources that will de-anonymize it; *"keep our promises regarding anonymous data, easy to de-anonymize by connecting the wrong data sources"* and *"we cannot control what happens with the data and how it is used"*. Even though users sign a contract and answer questions, they often ask for *"support from the providers"*.

Before a new data source is added to DDM, the external user has to answer multiple questions, *"review of new requests; where does it come from? how will it be used?"*. Compared to another team that handles more sensitive data, such as pictures and videos, the process becomes more complicated. They have to have well-defined use-cases, to *"make sure that we have clear use-cases"* and then they also have to notify the customer on what they are going to do with the data, i.e. they are *"entitled to let the customer know what the data is used for"*.

Governance

From the results in figure 4.1 we get that *processes*, *communication* and *discoverability* were the final themes for *acquisition*.

Participants detailed that the acquisition of data within the company primarily revolved around communication, processes, and use cases. Communication was especially a standard description of the initial phases of acquiring data. Among the actors acquiring existing data,

as the data user and re-user, a mixture of knowing about the existence and contacting colleagues close to the data was detailed to initiate access. One participant explained it as *"I would talk to people that I know is in close proximity to the data"*. Detailing specific tools or marketplaces for the discovery of data was not done by either of the users, more so the discovery through communication, as stated by the end-user: *"I was introduced through a meeting"*.

Further, acquisition seemed to involve following a pre-defined process that addresses access control and usage contract. Accessing this information was not as straightforward and often depended on where to find the information. For example, a producer within department 1 explained it as *"If they have not heard about it, they need access and contact me for further information [...] not everyone figures out that you should go through DDM Confluence site to find the information"*, where Confluence refers to their platform for documentation. For the actors situated in a provider and producer role, we saw that most acquisitions is handled through continuous communication, where communication involved reviews of requests or direct contact.

When looking toward future acquisition, participants focused on the importance of making it easier to discover the data and handling access control, and defining use cases. To work toward acquisition about use case was mentioned by one participant to be the biggest challenge: *"You have to get reasonable data to use-cases, there you have to find a collaboration [...] To get the use case right is the challenge"*

Relationships had a lot of different limitations, the concluded themes from figure 4.1 were *roles*, *communication*, *responsibility* and *feedback*.

A theme we saw from the participants was that relationships revolved around the coherent understanding of current roles and responsibilities and communication between these parties. All participants described forms of relationships with other actors. Unsurprisingly, the relationship most occurring was between the provider and consumer of the data, whether it be an individual or another department. In these relationships, participants described that communication was a key aspect and the most limiting factor today. To maintain a good relationship, the participating providers and producers stressed the need for feedback. A participant explained it as *"There is not a lot of feedback, you sometimes hear of users having trouble writing plug-ins [...] You become home-blind"*, *"Happy to have the feedback-loop short."*

Relationships also depend on a shared understanding of roles and responsibilities. Especially apparent was when discussing collaboration between the departments, as one keystone put it: *"Every team varies and should have their data ambassador"*.

From the results of *competition* in figure 4.1, *knowledge* was concluded as the singular theme.

The competition was not as prevailing as other topics throughout the interviews. Most participants responded that they were unaware of any existing competition. Naturally, as all actors work within the same organization, competition in the sense of competing for value was not seen. On the other hand, an indication of competition around resources and funding was noted. One participant stated that *"we are a new team and have to show our value as a team [...] [There is a] temptation to do advanced stuff just to show what we can do, instead we should explore the potential in the more basic stuff"*. If looking at the competition in ways of the best working with their data, more positive effects were detailed. By joining a Data Ecosystem, effectively open up transparency toward their silos, the benefit to gain knowledge from each other was discussed: *"It could be good to see how and what data other departments are collecting to*

see similarities and differences, to be able to improve our way of working".

Evolution

For Evolution, we had one sub-category; *Maturity*.

Maturity resulted in a more diverse set of keywords, although all revolving around features. From this the concluded themes, as seen in figure 4.1, were; *knowledge*, *terminology*, *discoverability* and *overview*.

The participants' prominent focus from participating in an ecosystem was increasing knowledge and unifying terminology. The interviewees felt that they do not have a common language when it comes to collaborating around data; *"how we talk about data, data sharing, how we meet between the silos"*. Some participants even expressed that, due to the pure technical knowledge needed, it is hard to collaborate, e.g., *"technical debt that makes it harder for not-as-technical to understand if they have access to the data, etc."* or *"stakeholders have a notion what they want to do, but not a clear objective in what they can ask about in regards to the data and what can be done"*. The technicality of data was something seen as a problem and something that wanted to be mitigated, were they expressed: *"we work toward making it possible, in a centralized way, to use the common data-lake, with pre-requisite to not making it difficult for each other"*. One interviewee even meant that before collaborating between departments, they have to mature, *"we need to mature even more before having standardized ways of working with data"*

A common issue was a lack of discoverability and general metrics around the usage of the data within the company. This issue could be divided into two areas; the producers and providers wanting more insight into whom is doing similar work and, secondly, to get more feedback from the users, expressed as; *"just how they use the data and if they are using it"* and *"to be able to see who uses the data would be useful"*. Another evolutionary aspect was related to the increased transparency of the data was to easier see what data is used, *"There is ongoing work to be able to see what data is and is not used so that unused data can be removed quickly to be able to reduce cost and resources"*, However, also to be able to give feedback to the users that they are not using what was requested, as depicted by a producer as: *"Be able to provide feedback to their managers that they are not using it"*. This while the users and re-user wanted to be able to see what databases that exist within Axis, *"developers spend 10-20% developing... the rest of the time is to hunt for data, figure out what is needed, talk to stakeholders to get definitions"*

4.2 Solution Design

For our solution design the goal was to, based on our conceptualization, synthesize a set of requirements toward a technical platform in the realm of data sharing. As a result of the synthesis, 14 requirements were discovered toward a technical platform.

The following section is structured into sections following the conceptual model, where our findings from both the case study and literature study are summarized under thematic areas followed by a presentation of the resulting requirements.

4.2.1 Value

When it comes to the value of data, findings from the interviews were primarily the increased knowledge and insights for the actors. Knowledge was referred to as personal understanding and decisions related to product improvements or features. Further, the value of knowledge seemed to be highly coupled to collaboration, effectively the value of collaboration. An important and existing collaborative relationship we noted was the one between the provider of data and the user. Collaboration between the different departments providing data was not formally established and prone to challenges.

Challenges around collaboration
The interviews further depict collaboration around data as a challenging area, both in the present and future. Summarized into three, the overarching challenges we found were:

F.1.1 The level of maturity varies in regards to understanding and working with data.
F.1.2 Lack of insights and standardized ways of working with data, especially between the departments providing data.
F.1.3 Lack of feedback in the current collaborative relationships.

For challenges around collaboration, both Nahar et al. and Kim et al. found challenges to be connected to the interdisciplinary nature of collaboration around data [23, 29]. Examples detailed where incoherent understanding and knowledge directly affected communication, documentation, and goals [23], also a lack of standardizations around tools and processes was noted to challenge efficient collaboration directly [29]. The suggestion here, from both mentioned papers and literature around Data Ecosystems, is the importance of working toward creating a coherent and standardized environment for collaboration [17, 23, 29]. Further, establishing clear and formal goals around the collaboration, using a common language and terminology, was also noted as a suggestion for improvements [29].

We see the potential of introducing a common platform to address the collaborative challenges faced at the case company and literature, where the potential resides in the fact that the platform provides a common terminology and standardized arena for the collaborative tasks moreover provides a transparent landscape around the use of data. From this, we suggest the following requirements:

Req. 1 *The platform shall support and present a standardized environment around its components.*
Addressing **F.1.1** and **F.1.2** above; the standardization of the platform around terminology, e.g., roles, relationships, data intrinsics and actions may help the actors to conform to a coherent understanding of the landscape.

Req. 2 *The platform shall support transparency around who uses the data and for what.*
Addressing **F.1.2** and **F.1.3** above; transparency may support insights around how the data is used and presented, further providing feedback between the involved parties.

4.2.2 Intrinsics

As for the intrinsics of data, i.e., regarding the types, quality, and legal aspects, our findings from the interviews were shown to be quite varying. Nevertheless, synthesis toward a technical platform, a couple of common directions were discovered of which proposition of

requirements were done. These were; Types of data and interoperability, legality and combination of data, and; Handling the quality of data.

Types of data and interoperability
From the interviews, we recorded that the actors within the case company collect and work with varying types of data, both regarding the characteristics and environment around the data, across the departments. Further, we discovered that each department was quite specialized in handling their specific types of data, e.g., amount, frequency, and protocols for acquiring new data. For example, two department specialized more on machine-generated data, as telemetry, log data, image- and video data, while the third department more on human-constructed data as e.g., forms and documents - all varying in how their environment was arranged, amount of data existent, frequency of collection, and sensitivity of the data. However, one commonality marked among the actors is the one belonging to the same organization, where data is of value and where all data in some sense can be classified as business-data, implying sensitivity to some extent. As for the challenges observed around the types of data, we concluded them to be:

F.2.1 Interoperability becomes a key challenge due to varying types of data and ways of acquisition.
F.2.2 Data within the case company is sensitive but may further vary between the departments depending on the characteristics.

Findings from the literature related to the characteristics of the data types highlight interoperability and sensitivity to be important areas to address for efficient data collaboration in a Data Ecosystem [17, 22], further possibilities and willingness to share [35]. These directives might pose a challenge for the case company in the current environment. First, as interoperability is crucial for the potential of combining data, which is also an important value driver in a Data Ecosystem [35], our recommendation is to emphasize the need for standardizations around the data before sharing, more specifically a common exchange reference model. Further, addressing the sensitivity and openness as a factor in deciding what can be shared, we recommend introducing and establishing a general framework for accessing and classifying the aspects of the data, e.g., as proposed by Enders et al.[14].

Transferring these recommendations (F.2.2 will be addressed later) toward a technical platform, we derived the following requirement:

Req. 3 *The platform shall support a standardized exchange reference model for different types of data.*
Addressing **F.2.1** above; by supporting a common exchange reference model applicable for different types of data, support for interoperability between different data types may increase.

Legality and the combination of data
Highly affected and coupled with the intrinsics of the data, more so related to the sensitivity of data, is the legality of sharing data. From the case study, it became apparent that the utmost important factor for the case company's data was that it should stay anonymous if anonymous. Anonymous data, i.e., data not containing personal information or other human-related data, did not have any legal constraints as long as it stayed within the company. However, as discovered in the literature, if combining data with non-anonymous data,

i.e., by joining the two data assets, the result would be to de-anonymize the data [22]. We observed that simultaneously, as the combination of data was referred to as a key factor to potentially increase the value of data, it was also noted as the most endangering act toward data anonymization, primarily as it is hard to monitor. Currently, to combat and mitigate the risk, access control in the form of regulatory contracts was in place, where responsibility and compliance of usage to keep the data anonymous were stipulated, as well pre-defined use-cases.

To summarize, the challenges and current recommendations observed within the case company were:

F.2.3 Anonymous data needs to stay anonymous.
F.2.4 Access control and regulatory contracts exist and are in use.
F.2.5 Combination of data is difficult to monitor, but may be very valuable.

Despite the literature primarily focusing on legality in an inter-organizational context, we discovered similarities between our findings.

A general lack of existing legal frameworks around the data in regard to ownership and responsibilities, as correspondingly requested in the evolution of Open Data Ecosystems [35]. As for the challenges related to non-anonymous data, literature provides a similar view on difficulty around sharing sensitive data, i.e., human data, being under regulations as the EU personal protection directive (GDPR), which addresses and enforces data minimization and limits use [28]. To still be able to increase the transparency of which data exist, without open access to the data, findings from the life sciences handling sensitive data suggest the use of metadata as an instrument for the assessment of the potential of reuse based on analysis [7]. However, as enforced by GDPR, the minimization of use still poses the challenge of incorporating clear use cases, which nonetheless affects the use of metadata.

Finally, as for the challenge of monitoring the evolution of data, i.e., combining and reusing data, relating findings from the literature was the emphasis on the need for validating the data against its intended usage by the providers, further the ability to monitor the data sources and usage of data, detailed as an essential aspect of an Open Data Ecosystem [21].

Concluding our findings from our synthesis, we propose the following triplet of requirements:

Req. 4 *The platform shall support the extraction of meta-data from the dataset.*
Req. 5 *The platform shall support the presentation of meta-data around the dataset.*
Addressing **F.2.2** from previous section and **F.2.3** above; to allow for the assessment of the dataset anonymously, allowing further contracts if the data is wanted.

Req. 6 *The platform shall support monitoring of the evolution of the data.*
Addressing **F.2.3** and **F.2.5** above; where the evolution of data includes, but is not limited to: The usage, classification, linage and, transformation as updates and changes of the dataset.

Handling the quality of data
Finally, as highly coupled to the intrinsics of the data, the critical aspect of data quality was highlighted in the literature as for sharing in a collaborative environment, and similarly observed in the case study.

From the interviews, we noticed the quality of data to be of concern for all participants, as connected to the potential value of data. However, similarly, as for the types of data, the handling and requirements concerning quality also varied between the departments. Recordings of more general quality concerns, e.g., outliers, missing values, and completeness of the data, were detailed to be handled directly by the producers, e.g., through statistical analysis. As for more use-case-specific quality concerns, we noted the responsibility to shift toward the data users, i.e., end and re-users, and addressed and handled through communication back to the producers and providers.

Our observations suggest that for data quality, the current challenges and recommendations at the case company were:

F.2.6 Handling quality concerns through statistical analysis is done to increase the quality.
F.2.7 Use-case-specific quality requirements are set by the user and handled through communication.

The importance of addressing data quality is something apparent in the literature, cited as a crucial obstacle to address for efficient data utilization, both in an open and closed setting, as well as in deciding its openness [21, 23, 14]. Generally, one challenging area considering the quality of data is that it cannot be judged without considering the context or situation, e.g., type of data, what the intended use and importance [21]. In the context of exchange in a Data Ecosystem in regards to handling the quality, suggestions from the literature are the establishment of agreements on quality standards once collaboration around data becomes active [21, 17]. For example, the use of Service License Agreements (SLA) between user and provider may help to ensure quality in its context, or even a quality management system [21, 44]. Further, to keep track and monitor quality attributes as adaptable to its use cases is an integral part of value and exchange of data, as the use case provides context [21].

Transferring our findings toward a technical platform, we suggest the following requirements:

Req. 7 *The platform shall support communication around the usage of data.*
Addressing **F.2.7** above; the platform should include communication around the data in order to assess the quality in relation to its context of usage.

Req. 8 *The platform shall support statistical quality checks of the data.*
Addressing **F.2.6** above; in order to support the current way of working, the platform should address the existence of quality checks as missing values or outliers.

Req. 9 *The platform shall support the linkage of documents between actors and related to the data.*
Addressing our findings in the literature regarding the existence of standardized quality agreement around the data quality, i.e., such as SLA's. Further also supporting the requirement **F.2.4** from **Legality and the combination of data**, which entails the existence of contracts.

4.2.3 Governance

For governance, the overall findings from the case study were that for data acquisition, relationships and competition between the actors mapped to themes regarding processes, com-

munication, roles, responsibility, and increased discoverability. Across all areas mentioned above, i.e., acquisition, relationships, and competition, we observed that knowledge and communication were an active and reoccurring part of the current and future governance structure, which resulted in the final areas of interest to be knowledge and relationships.

Acquisition through knowledge and communication
At the case company, for governance concerning the acquisition of data sources, emphasis on communication was witnessed as a vital part of the initiation of acquiring data, also further along the process. Examples of governance practices mainly done through communication were the discovery of data sources and guidance toward the subsequent processes around acquiring the data. We observed that all departments appear to have some sort of pre-defined governance protocol to acquire new and existing data sources.

Examples were the existing contracts for establishing use cases and defining regulations and responsibilities, although also observed not necessarily being easy to find, referring to communication. Further, the tribal and general lack of unified knowledge about where to discover the data was observed as a hinder today concerning the acquisition. Following are the summary of our findings in regards to present challenges and recommendations:

F.3.1 Communication was noted as a present and vital part of data acquisition.
F.3.2 Access control procedures are in place, although not necessarily easy to find.
F.3.3 The discovery of data was highly connected to existing knowledge and done through communication, also depicted as a challenging area today.

An important aspect raised in the literature regarding governance in a Data Ecosystem is the importance of obtaining information about the licenses and contacts for the actors [21]. The discovery of data, imagined to be an essential part of a collaborative environment around data, should also be supported in a unified way to establish knowledge around which data exists, who is responsible, and further information on how to acquire it. Being able to search, find, request access can consequently be seen as essential features in a Data Ecosystem [21].

Concluding our findings at the case company, in relation to the literature, we suggest the following requirements toward a technical platform:

Req. 10 *The platform shall support the presentation of documentation around the data.*
Addressing **F.3.1** and **F.3.2** above; for the acquisition of data, information around licenses, contracts, and other processes and information of whom to contact should be provided to establish knowledge of the data and acquisition process.

Req. 11 *The platform shall support the discovery of existing resources.*
Addressing **F.3.3** above; data sources within the ecosystem should be searchable, findable. and request access is an essential part of a Data Ecosystem and should be handled by the technical platform.

Relationships
For relationships in governance, our findings from the case study detailed the existence of current relationships, which also faced some challenges.

As suggested by the literature, in an ecosystem with actors of varying roles and responsibilities, a common understanding of the roles and sharing of these are important [27]. Correspondingly, as noted as a challenge at the case company, the suggestion of each department

having a data ambassador was observed from our findings. The case study also highlighted that feedback and communication between the different actors with different roles in the existing relationships was essential but also problematic. Overall, the challenges found categorized into:

F.3.4 Relationships were highly related to responsibility, where a common understanding of the roles is important.
F.3.5 Communication, and especially feedback, was seen as an essential part in relationships, and also troublesome.

From the literature, the importance of actors to understand the roles is detailed as an example of significance, as governance mechanisms may vary across the Data Ecosystem and roles are connected to responsibilities, especially in an inter-organizational context [27]. As found for efficient collaboration, which is of direct value in a Data Ecosystem, similarly a common understanding as a suggestion should be applicable for the relationships in a Data Ecosystem [17, 29, 23].

The question and importance, of trust is further emphasized in the literature regarding the relationships in a Data Ecosystem, which we did not observe at the case company. The importance of trust, enabled by transparency among the actors, was described as a critical factor for building and maintaining good relationships [27, 17]. The reason to why trust was not found in the case study may be related to being in an intra-organizational context, which differs from the literature. However, as actors still exchange resources connected to responsibility, we suggest that building trust and transparency among actors should be considered a vital element of relationships.

Finally, as detected in our case study, communication and feedback were also topics depicted in our literature findings. Here, the existence of a discussion and feedback of usage was seen as an essential element in a Data Ecosystem, including but not limited to the relationship [44]. E.g., the possibility to give feedback about certain particularities or ask questions about the data could be beneficial for other actors as well.

Concluding requirements based on our findings, we suggest the following:

Req. 12 *The platform shall base its governing mechanisms on established, common roles.*
Addressing **F.3.4** above; if new governance mechanisms, e.g., regarding access rights and responsibilities, are introduced or transferred onto the platform, these should be based on the ecosystem's common roles due to the importance of a common understanding.

Req. 13 *The platform shall support feedback and discussion around the usage of the data sources.*
Addressing **F.3.5** above; as communication and feedback are important parts of relationships, the platform should support the possibility for feedback and discussion around the data.

Req. 14 *The platform shall support transparency around the provision and usage of data.*
Addressing the importance of trust in relationships between the actor, concerning our findings from the literature.

4.2.4 Evolution

As for evolution, direct comparison with the literature is challenging due to the novelty of Data Ecosystems. Moreover, as the case company has yet initialized a Data Ecosystem, the recordings of evolution and maturing were discovered to provide less concrete findings.

Nevertheless, results from the case study highlighted an interest in a Data Ecosystem among the actors, with primarily the motivation of the value of collaboration. Initiating collaboration in a Data Ecosystem on a common, underpinning technical platform, i.e., platform-centric Data Ecosystem, was seen to mature as an organization, increasing knowledge and unifying terminology, ways of working, and processes. Comparing this to our findings regarding the maturity of Data Ecosystem in literature, we also observed the importance of standardizing the ecosystem around a common platform [35].

Further, as not discovered in the case study, the importance of business models to be incorporated in the Data Ecosystem is brought up as an important evolutionary step. Our findings from the literature also note the importance of the participants have to work toward enabling a common *"data mindset"*, as the collaborative data environment evolves [25]. Finally, as for evolution, the need for tools to support Data Ecosystems and enable data sharing should be developed and standardized [35], as, e.g., can be seen as a significant contribution in making OSS widely adopted [32].

4.3 Validation

Validation of the proposed solution, i.e., the requirements towards a technical platform, was performed by the two activities of surveying the case study participants through a questionnaire and conducting an assessment of existing platform solutions. This section first details the results recorded from the survey, followed by the resulting assessment of existing platforms.

4.3.1 Survey of Requirements

After distribution, the questionnaire was left open and available for roughly five days in order to get as many participants as possible to respond. For the final questionnaire, see Appendix D. After five days, 8 out of 9 approached participants had responded on the questionnaire, resulting in a response rate of 89%. The concluding results are presented in figure 4.2 and for more details see Appendix E.

From the figure 4.2 displaying the recorded data from the questionnaire, we can see that the most agreed requirements, on average, were 1, 2 and 14. Four responses detailed disagreement of a requirement, but each one where on different requirements; one of them was even on a requirement with higher positive opinions, requirement 2. There were only two requirements for which all responses were above neutral; 3 and 14. Requirement 9 had the lowest score, neutral, with most of the responses being neutral and agree. On average, the participants had a positive opinion toward the requirements.

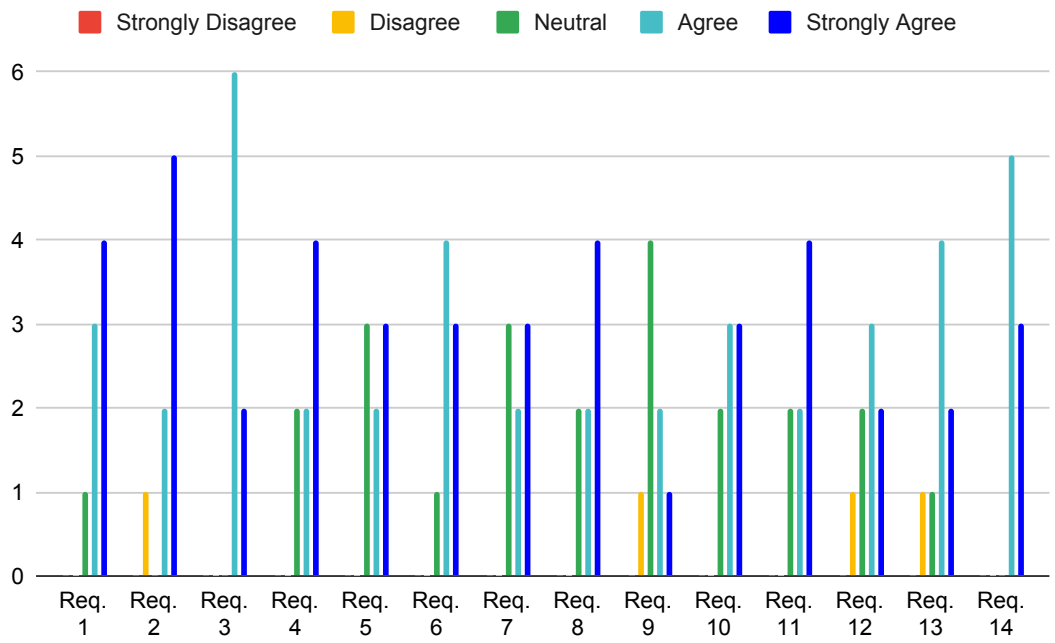


Figure 4.2: Recorded results from the validation questionnaire.

In the questionnaire we included two optional free text questions at the end, see Appendix D. We received three open responses: The first was one emphasizing the suggestion of *"Clear owner and transparent responsibility of different roles"*. Further, another one was detailing an interesting discussion point around the cost in the future, i.e., the actual cost of the future implementation and maintenance of such a platform; *"All of these requirements will, in practice, add a cost. It takes time to find and establish common roles, update and unify meta-data, find owners etc."*.

The final comments were direct feedback toward the questionnaire, where the participant suggested establishing more explicit use cases around each requirement. Even though not directly applicable due to the objective of the survey, i.e., record the actors opinions around the requirements, the feedback was valuable and noted as future improvements.

4.3.2 Platform Assessment

Here we are to compare the given data catalog platforms toward requirements discovered above. Only the features and documentation given on their websites [1, 4, 36], accessed *2022-01-15*, were used for the comparison with the requirements in table 4.1.

	Datahub	Amundsen	Secoda
Req. 1	✓	✓	✓
Req. 2	✓	✗	✓
Req. 3	✗	✗	✓
Req. 4	✓	✓	✓
Req. 5	✓	✓	✓
Req. 6	✓	✗	✗
Req. 7	✗	✗	✓
Req. 8	✓	✗	✗
Req. 9	✓	✓	✓
Req. 10	✓	✓	✓
Req. 11	✓	✓	✓
Req. 12	✓	✗	✓
Req. 13	✗	✗	✓
Req. 14	✓	✓	✓

Table 4.1: Technical Platforms checked against the requirements.

Req. 1 - All of the platforms fulfilled this requirement by using one common interface for all connected data sources.

Req. 2 - Datahub and Secoda fulfilled this requirement, but Secoda went beyond what Datahub had. Datahub mainly did this by displaying who the users were, and then how the data was used could partly be viewed with the lineage tool and, in some cases, see queries. Secoda also had this, but beyond that, they solved it by users being able to document how they used the data and its results within the platform. Amundsen failed this requirement; they only displayed the potential users of the data and nothing else.

Req. 3 - Only Secoda managed to fulfil this requirement. Secoda could do this by having a simple *Data Request* feature; it was easy to access data and check requirements with this feature. Both Datahub and Amundsen were missing something to support this.

Req. 4 & 5 - All the platforms supported these two requirements. With Datahub and Amundsen, you had to write code to be able to extract data and present it on the platforms. The commercial and more user-friendly platform Secoda solved this without having to write a line of code. See section 2.3 for more information.

Req. 6 - Datahub was the only platform being able to support this request. With the feature of *data lineage*, historical statistics with were and changes, they could solve this requirement. The other two were missing features for this.

Req. 7 - Secoda was the only platform with integrated communication around the data. It also made it possible to talk about data in third-party apps with Secoda integrated into them as support. In Datahub, you could see the team who owned the data and then get their email or similar, but this was not enough to fulfil this requirement.

Req. 8 - Amundsen had a few fundamental static quality checks but not enough to check the mark. In Datahub, you could get detailed statistics and quality checks of the data and each row of the data. From what we could see, Secoda did not have anything supporting this

requirement.

Req. 9 - All the platforms solve this requirement mainly with the tagging system and documentation feature on the source.

Req. 10 - All the platforms supported this. However, with Datahub and Secoda, you could add more specific documentation, queries, and documentation for each column in the data source, e.g. name, id.

Req. 11 - All the platforms supported search in similar ways, search for tags, teams, sources, etc.

Req. 12 - With Datahub and Secoda, you could set custom access rights to different sources based on easy to create roles/policies for teams and users. This was not possible in Amundsen.

Req. 13 - Only Secoda fulfilled this requirement, with the ability to connect the platform to third-party apps and the two features called *Data dictionary* and *Data request*.

Req. 14 - By being able to set owners and users on the sources, all the platforms managed to fulfil this requirement.

As we can see in table 4.1, Datahub and Secoda fulfilled most of the requirements, with Secoda only missing two and Datahub three. Amundsen, however, was missing seven out of the 14 requirements making it fulfil only half of the requirements. From our results, we see that Secoda was the only platform to fulfill requirements 3, 7, and 13, all including communication aspects around the data. In contrast, Datahub is the only platform to fulfill requirements 6 and 8, depicting more technical aspects. This result might indicate that Secoda emphasizes the communication aspects of working with data, while Datahub, on the other hand, seems to handle more technical aspects such as data linage, etc.

Chapter 5

Discussion

In this master's thesis, we conducted a literature study to find challenges and benefits in the literature, further, a case study at the case company, see section 4.1. Through a synthesis of the gathered material, we identified 14 requirements for a technical platform in order to mitigate the found challenges, see section 4.2.

This chapter begins with a general discussion of our results, followed by a discussion of the chosen methodology. Lastly, a discussion of potential future work is presented.

5.1 General Discussion

Returning to the initiation of our research, the overarching goal of our work, and the motivation behind, was to investigate the benefits and challenges concerning data sharing in the emerging concept of Data Ecosystems. The ambition was to review the current literature around Data Ecosystems, further investigate the concept in an intra-organizational context. As a final contribution, we also wanted to anchor the research to the practical context by proposing a set of requirements to consider for a technical platform supporting such an ecosystem. Given our goals with the research, the first point of discussion is naturally to which extent we consider that our goals have been reached. Our research resulted in the solution constructs as planned, although as in any work, our findings do entail some discussion points.

One aspect affecting the results, which was a recurring point of discussion throughout our research endeavor, was ***the maturity of Data Ecosystems***. As the concept of Data Ecosystems is still in its infancy, only recently emerging in the literature, the current research around the topic is relatively sparse and, on some level, incoherent as detailed in section 2.2. An effect

which we noticed in our research was primarily concerning investigating RQ1 and in extent RQ2.

As the aim was to investigate the benefits and challenges around Data Ecosystems from varying directions, the novelty of the research area resulted in the literature, further interpretation of Data Ecosystems at the case company, becoming quite sprawling. If only looking at the resulting artifact from the literature study, our findings can be concluded to only partly reside in research around Data Ecosystems, but to a large extent also in neighboring research domains. Similarly, as for Data Ecosystems, the value of collaboration and challenges faced were discovered in Machine Learning (ML) research, data governance components in research of Data Management, and the importance of data intrinsics as a frequent area of discussion in Big Data. On the other hand, the findings from neighboring research areas also provided some interesting results, which necessarily were not found in the current research of Data Ecosystems. For example, suggestions of solutions for the challenges in collaboration found in ML and Data Science research, or ways of working with and sharing sensitive data are more frequently discussed in the life sciences. As our research resides within an emerging concept in the literature, naturally incorporating findings from neighboring research domains is an entirely acceptable and vital aspect of our research.

The mixture in our literature findings however opens up the question of whether a Data Ecosystem was the best and most suitable concept for our investigation into the challenges and benefits faced at the case company, with regard to collaborating around data. For example, the ML research provided another explanation of the landscape as one of ***interdisciplinary collaboration***, decoupled from the concept of an ecosystem. This research instead provided more focus on processes and best practices for data sharing between departments, and the organizational aspects related to this [29], which in some sense corresponds better practical context at Axis. However, we would still argue that even though our participants worked at different departments with different roles and tools, the distinct interdisciplinary aspect, e.g., the one between Data and Software engineers were not found to a great extent, nor focused on. This might have become more prevailing if we would have focused on later aspects of data sharing, e.g., value generation through ML. The departments we investigated instead all generated their unique value from their data, also doing this quite autonomous and specialized as of today. On the other hand, this autonomy does not correspond well with the ecosystem metaphor which again raises the question there are better ways to describe the collaborative environment at the case company. However, the motivation behind our research was partly the increased interest in collaborating on their data within the company which may, in the future, involve more interactions and positive value-generating dependencies which are not present as of today.

On the same topic, where the biological meaning of an ecosystem can be argued to already exist can be argued to be in the relation between the provider and user of data, where they in some sense depend on each other to survive and thrive. Especially from DDM's perspective, their value was discovered to in great extent depending on first being generated by the users or analytics by a re-user on their data. For example, the value of the data is better and improve their products, or locate and fix existing bugs, etc., which is first done by the users of the data.

The novelty and potential mismatch of the concept Data Ecosystem further displayed itself in another way at the case company, as the discussion of ***varying maturity of the participants***.

Even though no current Data Ecosystem exists at the case company, the idea of working in a collaborative environment around data and terminology of actors and relationships was not as alien as we first thought. Still, the resulting artifact of the perceived benefits and challenges was difficult to map toward a coherent understanding of a Data Ecosystem since the degree of common understanding of the concept was quite low and hard to coherently install. An example of where this became apparent was concerning the evolution, where, e.g., one participant discussed the potential of the Data Ecosystem in the forms of features, while another more reflected around the next steps in the current landscape and maturity for the department.

As discovered primarily through the synthesis toward investigating RQ3, *"What requirements should a technical platform fulfil to address the challenges found in RQ1 and RQ2?"* our results from the synthesis of the two studies still showed to share quite a lot of similar themes regarding benefits and challenges. Deciding to follow an existing conceptual model, in our case the one of an Open Data Ecosystem by Value, Intrinsics, Governance, and Evolution could be argued to be both beneficial and restraining. As for benefits, the conceptual model helped us fixate and scope our research navigate the novel landscape of Data Ecosystems. Further it helped us to connect our findings from neighboring research areas, and ease our synthesis between the data recorded in the two studies. Another benefit was that the research leading up to the conceptual model provided us with background research around the concept and initial findings. On the other hand, the conceptual model was one of inter-organizational context, further encompassing open data, which is not matching the practical context of the problem instance we set out to investigate. Here, the question of how well the conceptual model of an ODE suited our context may be discussed. Primarily, the aspects for ODEs of need for business models, licenses around the data, and handling of the competition were not as prevailing at the case company. However, even though much of the literature on Data Ecosystems enfolded an inter-collaboration context, many aspects of an inter-organizational data collaboration were still quite applicable in the intra-organizational context investigated, e.g., the value of collaboration, importance to address the intrinsics of data and support relationships and transparency.

The most significant difference between our findings in the two studies was the aspects of trust and competition among the actors. We did not record any implication among the participants in the case study regarding trust, and competition was less prominent than in the literature. The feasible explanation for this could be that trust and competition are naturally established and less prominent simply by belonging to the same organization, i.e., there is no direct competition among the actors working toward and taking part in the same success as the company's entity. However, some observations from the case study told another story. As reported in our finding, one participant provided the notion of resources and funding toward the department to be connected to the generated value of data, which opens up the much important question of existing competition within an organization. Comparing this to the literature of inter-organizational context, competition among the actors can be a crucial factor towards the willingness, and risk, to share data with other actors. This notion might suggest that, even for an intra-organizational context, competition may affect the willingness to collaborate and share data and should therefore be taken into consideration. The participating departments investigated in our case study all handled a quite diverse set of data types, further providing unique value to the company with regard to each other. This might be why

we did not record any further elaborations regarding the topic of competition. However, we see and highly encourage the aspect of ***inter and intra-organizational Data Ecosystem*** to be further explored and discussed in correspondence in future research around the concept of Data Ecosystems, especially the similarities and differences.

In our case study, the topic of communication was discovered to appear frequent among the participants. This was especially true regarding governance, e.g., in discovering and acquiring data, directing to processes, and a cornerstone in current relationships. In contrast to the literature on Data Ecosystems, we did not discover communication as a pivotal area of consideration. However, as for the challenges in collaborating on data, especially in an inter-disciplinary context, communication was frequently depicted. If communication is an effect of being in an intra-organizational context, or an effect of other factors is discussable. Belonging to the same organization and being colleagues with established communication channels, such as meetings and social platforms, might explain the more frequent appearance in the case study. On the other hand, we also saw communication as a lack of feedback between actors and a hurdle in discovering data. As also found, in conjunction with the ***need for internal maturity***, communication was also seen to effectively improve in connection to maturing, e.g., in knowledge, transparency, and standardized ways of working.

If just looking at these findings from the case study, alterations of the conceptual model of ODE we followed should be considered. As we discovered, aspects of governance and evolution were hard to map towards our findings in the intra-organizational context. Here, for example, more of the cultural aspects, and challenges, related to the motivation of maturing as departments and organizations as behaviors and processes working with data, could be argued as an important aspect in the evolution. Further, with an emphasis on the social aspect of a Data Ecosystem, we see the importance of communication, both to be included and improved, as a vital aspect to consider moving forward with the idea of enabling an intra-organizational Data Ecosystem, and should therefore be encompassed in the conceptual model. For example as an addition as an integral part in regard to the value of data and collaboration and governance, similarly as described in the interdisciplinary context of ML-research [29, 23]. However, we would still argue that aspects from the openness of ODE as handling competition, potentially internal business-models, but also regarding cultural aspects as more open and coherent processes and behaviors to be fruitful, and if not in the current, potentially in the future.

Transferring this further toward the potential of introducing a common platform, working with the data in a coherent environment, providing more transparency, and a standardized way of working can potentially help to improve both maturing and communication around the data. On the other hand, simply being a platform with its constraints it might be of no use if the actors are not yet mature, as also mentioned by one participant. Instead, focusing on maturing internally and around the concept of a Data Ecosystem might be the way forward before enabling a common platform.

Finally, moving toward our investigation and research around RQ3, our findings for discussion reside in both the solution design of the requirements and the evaluation of these. Overall, we would argue that for RQ3, and regarding the initial problem instance, the resulting artifact of requirements is a feasible solution moving toward the practical context.

As discovered in our survey, most participants agreed, i.e., had generally a positive opinion

toward the requirements, which is positive regarding the result being a feasible solution for the case company. Although, we consider that ***implementing the requirements*** is still to be done to further evaluate if they are feasible in its practical context. This, as noted by one of the participants, is potentially connected to a new set of challenges such as the cost related to the maintenance, further complying with the requirements in such an ecosystem platform.

As for the second part of the validation, i.e., assessing the requirements against existing data catalog platforms, even though brief, the analysis provides a positive indication of the feasibility of use for the requirements. We could see that the technical platforms of data catalogs today fulfill many of the requirements derived from our findings of the intra-organizational Data Ecosystem. It was further discovered that they sometimes fulfill the requirements very similarly, and other times in very different ways. Another aspect for discussion is the aspect of Open Source versus Enterprise, as Datahub and Amundsen are open source products, while Secoda is an enterprise solution. By fulfilling large portions of the requirements, further, as the platforms similar to our research have been developed to mitigate existing challenges around working and collaborating around data at the companies behind the platforms, we see the potential. We would like to recommend further investigating the potential of these data catalog platforms as a suggestion to both practitioners, at the case company Axis, and researchers as a part of the underlying technical platform for Data Ecosystems.

5.2 Methodology Discussion

Section 3.1 detailed that this master's thesis follows the design science paradigm with problem conceptualization, solution design, and validation activities. Generally, all the data collected and analyzed throughout our research was of qualitative nature, except for the data collected through our validation questionnaire. A discussion around both our methodological choices, implementation and effect of these in regard to the validity of our research is therefore of much importance.

As detailed in section 3.1, our research questions primarily resided in the problem and solution constructs emerging from the problem conceptualization and solution design phases following the design science paradigm. Motivation and description can be found in sections 3.2, more specifically 3.2.1 and 3.2.2, respectively.

Given the qualitative nature of the problem conceptualization, the internal, external validity further reliability was of great concern and actively discussed throughout our research. One decision, briefly discussed concerning our results in the previous section, was to follow a conceptual model of a Data Ecosystem. This decision meant for the internal validity, concerning both studies constituting the problem conceptualization, to provide a fixed scope for our research. In more practical phrasing, the model provided a scope of the areas to focus our investigation around.

Further, to some extent, the conceptual model was also of help to better construct our audit trail, i.e., anchoring our findings to the conceptual model instead of solely dependent on the mapping of findings by us researchers, which was imagined to potentially vary between the two studies. The ambition was to increase our research's internal validity and reliability, as transparency around our findings also became more explicit.

Although, as discovered, the scope provided from the conceptual model still resulted in quite a large amount of data, becoming extensive in relation to our time and efficiency constraints. Even though it is quite cumbersome to mitigate fully, it may be somewhat troubling regarding the internal validity. In retrospect, it might have been more suitable to select a smaller scope, e.g., focusing on a more isolated aspect of the problem instance or around an area of conceptualization.

Going further into the individual methods performed and starting with the literature study, we adopted a more systematic approach involving keyword search, mapping and analyzing the results. By following this, one ambition was to increase the internal validity given a more structured way of conducting a literature study. Still, as for any literature study, the selection of digital research libraries and keyword strings is known to affect the outcome. As for the digital research libraries, we opted to use the Google Scholar research portal, with the pre-defined decision only to include peer-reviewed articles accessible through Lund University Libraries. A threat of validity is the potential that other research libraries not included in our criteria may contain relevant literature research. However, we would still argue that opting for peer-reviewed literature, accessible through Lund University Libraries, which encompasses multiple research databases, was sufficient to increase the quality to a great extent.

In regard to the other impacting factor, i.e., constructing our keyword strings, we categorized ours with regard to the scope provided by the conceptual model; further, we did our best not to miss any potential areas of interest in formulating these keywords, still within our scope's relevance. We also discussed the keywords with our supervisors to get an outside opinion. Still, selecting keywords is a problematic step that highly affects the outcome, further hard to mitigate to the full extent.

As for the discussion of validity around our second activity, i.e., conducting semi-structured interviews, topics for discussion was primarily the effect of the selection of participants, the novelty of the research topic and generalizability.

Our case study was done through one-on-one semi-structured interviews, where the outcome of the case study was highly dependent on the selection of participants and further analysis. We chose our participants by a criterion sample around current expertise, roles and departmental belonging. Still, the selection was in some sense of convenience, being conducted within a single company and dependent on its departments and expertise. Therefore, we cannot assume that the results will generalize outside the company. However, as our questions stemmed from and later combined with the literature study, our findings were highly compared and connected to the more generalized research findings. As for any research, we promote replicating our study in different organizational contexts to research the topic further in order to build a better and more generalized empirical body of knowledge.

Furthermore, as investigating the novel topic of Data Ecosystems, we would collect opinions and hypotheses rather than facts and experiences. As for our benefits and challenges, the benefits were primarily affected by this, as discussed previously. However, potential misinterpretation during the interviews was still a concern as the knowledge around the concept was both at a novice level and varied between the participants and the authors. In order to mitigate this issue, before the interview, we provided background information on the topic at hand to the participants. For example, this information was an explanation of the motivation

behind our research, a presentation of Data Ecosystems and a conceptual model. Further, we also briefly introduced the Data Ecosystem concepts at the beginning of the interview session, intending to create a coherent understanding and clear any questions or misinterpretations.

Finally, regarding the solution design, i.e., synthesis of the previous findings toward the investigation of RQ3, the validity is again highly affected by the authors as analysts of qualitative data. Overall, one crucial area discussed before conducting our synthesis was establishing a process for maintaining transparency regarding the procedure and reporting of our findings. Examples were to plan for which and how the material was to be used in our reasoning, as other researchers might find other themes given their background and knowledge. We concluded to perform our synthesis in three iterations, one for each source of data. All the data involved in the synthesis was collected in our study and mapped to the same conceptual model, which hopefully increases validity, especially transparency. However, as the goal of the synthesis was the technical requirements, it may have impacted our analysis, i.e., if the goal would be different, the outcome of the synthesis would potentially also be different. Given this, we tried to provide the material leading up to each requirement.

As for external validity of the requirements, i.e., the relevance for other cases, we would still argue that the requirements being a product from both the literature and the case study, are still applicable toward a generic technical platform. However, we see the need for more research to explore similar topics, preferably at another case company, to extend and validate our findings.

The outcome of the synthesis, i.e., the product of requirements, was further validated through a questionnaire and analysis against three existing data catalog platforms. Given our problem construct to reside in the practical setting at the case company, and as the industry-research collaboration, we returned to the same setting for validation.

For the questionnaire, the biggest issue regarding validity was again the generalizability of the results. As the same set of participants was selected, the risk of biased answers and potentially case-specific opinions were seen as threats. However, as the technical platform is still to be implemented, the aim was primarily to see if the solution, i.e., requirements, was feasible and captured the problem instances found, of which the participants in the case study can be argued as most suitable to answer.

Another aspect of mitigating the risks of our validation becoming endangered regarding external validity, we decided to conduct the second part of the validation, i.e., the validation against the existing technological platforms of data catalogs. However, a quite limited scope had to be set primarily due to time and resources. Here, the biggest issue can be argued to be the pre-defined set of platforms, which the case company provided. To mitigate this issue, as there exist many platforms, also outside the realm of data catalogs, our suggestion would be to extend the scope and introduce a more rigorous procedure when selecting and evaluating candidates. However, as not captured by our research goal, we aimed to evaluate if the requirements are feasible for use and generalizability; we highly motivate and hope new research will investigate this topic further.

5.3 Future Work

Our master thesis has been conducted at the Axis Communication, moreover at the department of *DDM* which has naturally limited us to the type of data they have and the issues they face. As organizations may differ regarding their data, maturity, and issues they face, we recommend similar studies at other organizations and departments to further explore the topic.

Further, as the concept of *Data Ecosystems* is still relatively new and emerging in the literature, new and exciting research may appear; hence we see the need to revisit the literature for new findings and elaborations.

We discovered some differences between an inter-and intra-organizational setting for the Data Ecosystem. However, we believe that there is more to explore, as cultural challenges, and further is also the topic of Open/Private Data Ecosystems compared to Open/Inner Source Software Ecosystem, which we, unfortunately, did not have time to explore in our study, but find to be an exciting topic and would recommend for future research.

Finally, as for the requirements resulting from our synthesis, even though we hope that these can be used in future research, we still see a need for further validation, preferably at the company experiencing similar challenges. As for their practical application, we would have liked to evaluate these further in a practical setting, i.e., testing them against an implementation of a Data Catalog platform or extending outside our scope of Data Catalogs, investigate if there are other types of technical platforms supporting these better. However, our evaluation of the current Data Catalog platforms showed promising results, which motivate us to recommend future research to investigate the landscape of Data Catalogs for supporting the socio-technical concept of Data Ecosystems.

Chapter 6

Conclusions

Following the conceptual model of *Value*, *Intrinsics*, *Governance*, and *Evolution* as components of an Open Data Ecosystem, our research have explored and contributed to the existing body of knowledge around data ecosystems by applying the conceptual model in an intra-organizational and private setting, further populate it with our empirical findings. Our research also contributes with a synthesis of our empirical studies towards a set of technical platform requirements to support such an ecosystem.

For **RQ1** we conclude our findings to both reside in the literature of the emerging concept Data Ecosystems, but also in neighboring research areas. The challenges we found were connected to both the social and technical aspects of a data ecosystem, as examples of these are; the importance of assuring trust among actors, willingness to share and motivate collaboration further technical challenges of handling quality, interoperability of data, and lack of common processes and legal constraints. Interoperability, quality, and the sensitivity of data were especially challenging and significant obstacles for data utilization as they directly limit the potential of reuse, combination, willingness, and the possibility to share data in a collaborative environment.

For **RQ2** we found similar challenges as for **RQ1**, however with the distinction of shared knowledge and importance of communication to permeate many of our findings, also influential in the challenges and current practices of discovery, acquisition of data, and existing relationships. We did not find trust to be a challenge within the company as well.

As for general benefits of **RQ1** and **RQ2**, we conclude that working toward creating a coherent and standardized environment for collaboration, enabling clear and formal goals, and using a common language and terminology are of importance. We further see the benefit of having established roles and relationships, as it contributes to a more coherent environment, to be one supported by the concept of Data Ecosystem. Primarily from the literature, we

conclude that legal frameworks, business models, and standardization around technologies and processes concerning data sharing are necessary and must be developed in the future. These along with the need of assuring trust between actors as increasing transparency and equality around the data.

By introducing a Data Ecosystem for better discover, share and collaborating on data at the case company, our findings suggest the benefits are, among many, a way to increase insight and knowledge of the existing data, where the value of the data itself may increase and foster new innovation. The quality of the data can also improve by having a better and more efficient feedback loop and increased transparency between the actors. We also see the significant benefit of collaborating in a unified environment around the governance mechanisms as processes, roles, and responsibilities as well as the potential for the actors to more easily monitor and control who uses the data and for what. As for a common, underpinning technical platform, one of the largest benefit, and motivation, was found to be the ease discovering new data, decoupled from the current necessity of communication and knowledge.

Finally, as for the contribution of synthesis from the results from **RQ1** and **RQ2**, our investigation of **RQ3** concluded in a set of 14 proposed requirements toward a technical platform to support the Data Ecosystem investigated at the case company. The found requirements were as follows:

Req. 1 The platform shall support and present a standardized environment around its components.
Req. 2 The platform shall support transparency around who uses the data and for what.
Req. 3 The platform shall support a standardized exchange reference model for different types of data.
Req. 4 The platform shall support the extraction of meta-data from the dataset.
Req. 5 The platform shall support the presentation of meta-data around the dataset.
Req. 6 The platform shall support monitoring of the evolution of the data.
Req. 7 The platform shall support communication around the usage of data.
Req. 8 The platform shall support statistical quality checks of the data.
Req. 9 The platform shall support the linkage of documents between actors and related to the data.
Req. 10 The platform shall support the presentation of documentation around the data.
Req. 11 The platform shall support the discovery of existing resources.
Req. 12 The platform shall base its governing mechanisms on established, common roles.
Req. 13 The platform shall support feedback and discussion around the usage of the data sources.
Req. 14 The platform shall support transparency around the provision and usage of data.

Validation of these requirements showed promising results, both regarding opinions of their relevance among the actors further possible use for assessing existing platforms, i.e., Data Catalogs.

To summarize, we found that the concept of Data Ecosystems encapsulates many of the challenges found for the collaboration of data in an intra-organizational context, further providing benefits. We see the potential of introducing the concept of Data Ecosystem as a way forward to mature as an organization, increasing knowledge and unifying terminology, ways of working, and processes. However, as an emerging concept, we still see the need for fu-

ture research to delve deeper into the topic and monitor the evolution of the literature and neighboring areas, as for the case company to mature internally.

For the requirements toward a technical platform, we hope for the use of these at the case company moving toward investigating the potential of introducing a technical platform, e.g., a Data Catalog, to better discover, share and collaborate on data. We would also like to see future research validate our requirements further, also transferring this outside of the intra-organizational context.

References

[1] *A Metadata Platform for the Modern Data Stack | DataHub*. URL: https://datahubproject.io/ (visited on 01/03/2022).

[2] Rene Abraham, Johannes Schneider, and Jan vom Brocke. "Data governance: A conceptual framework, structured review, and research agenda". In: *International Journal of Information Management* 49 (2019), pp. 424–438. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2019.07.008.

[3] Joan E. van Aken. "Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules". In: *Journal of Management Studies* 41.2 (2004), pp. 219–246. DOI: 10.1111/j.1467-6486.2004.00430.x.

[4] *Amundsen, the leading open source data catalog*. URL: https://www.amundsen.io/ (visited on 01/03/2022).

[5] Charles Arthur and technology editor technology. "Tech giants may be huge, but nothing matches big data". In: *The Guardian* (Aug. 23, 2013). ISSN: 0261-3077. URL: https://www.theguardian.com/technology/2013/aug/23/tech-giants-data (visited on 01/15/2022).

[6] *Axis Communicaiton - About Axis*. URL: https://www.axis.com/en-ca/about-axis (visited on 10/08/2021).

[7] Martin Boeckhout, Gerhard A. Zielhuis, and Annelien L. Bredenoord. "The FAIR guiding principles for data stewardship: fair enough?" In: *European Journal of Human Genetics* 26.7 (July 1, 2018), pp. 931–936. ISSN: 1476-5438. DOI: 10.1038/s41431-018-0160-0.

[8] Maximilian Capraro and Dirk Riehle. "Inner Source Definition, Benefits, and Challenges." In: *ACM Computing Surveys (CSUR)* 49.4 (2016), pp. 1–36. ISSN: 0360-0300. DOI: 10.1145/2856821.

[9] Diane Coyle, Stephanie Diepeveen, Julia Wdowin, Lawrence Kay, and Jeni Tennison. *The value of data: summary report 2020*. Report. Bennett Institute for Public Policy, Feb. 26, 2020. URL: https://apo.org.au/node/277171.

[10] Daniela S. Cruzes, Tore Dybå, Per Runeson, and Martin Höst. "Case studies synthesis: a thematic, cross-case, and narrative synthesis worked example". In: *Empirical Software Engineering* 20.6 (Aug. 2014), pp. 1634–1665. DOI: 10.1007/s10664-014-9326-8.

[11] Bart Custers and Helena Uršič. "Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection". In: *International Data Privacy Law* 6.1 (Jan. 2016), pp. 4–15. ISSN: 2044-3994. DOI: 10.1093/idpl/ipv028.

[12] *Data*. URL: https://dictionary.cambridge.org/us/dictionary/english/data (visited on 01/10/2022).

[13] *Deductive and inductive approaches to coding*. Delve. URL: https://delvetool.com/blog/deductiveinductive (visited on 02/23/2022).

[14] Tobias Enders, Clemens Wolff, and Gerhard Satzger. "Knowing What to Share: Selective Revealing in Open Data." In: *ECIS*. 2020.

[15] *FAIR Principles*. GO FAIR. URL: https://www.go-fair.org/fair-principles/ (visited on 01/15/2022).

[16] Vahid Garousi, Dietmar Pfahl, João M Fernandes, Michael Felderer, Mika V Mäntylä, David Shepherd, Andrea Arcuri, Ahmet Coşkunçay, and Bedir Tekinerdogan. "Characterizing industry-academia collaborations in software engineering: evidence from 101 projects". In: *Empirical Software Engineering* 24.4 (2019), pp. 2540–2602. ISSN: 15737616. DOI: 10.1007/s10664-019-09711-y.

[17] Joshua Gelhaar and Boris Otto. "Challenges in the Emergence of Data Ecosystems". In: *PACIS 2020 Proceedings*. June 2020, p. 175. URL: https://aisel.aisnet.org/pacis2020/175.

[18] Ana Gillan. "Governance: the key driver for data-driven innovation". In: *Computer Fraud & Security* 2021.4 (2021), pp. 10–13. ISSN: 1361-3723. DOI: 10.1016/S1361-3723(21)00041-5.

[19] Martin Höst, Björn Regnell, and Per Runeson. *Att genomföra examensarbete*. Studentlitteratur, 2006. ISBN: 9789144005218.

[20] *How To Do Open, Axial, & Selective Coding in Grounded Theory*. Delve. URL: https://delvetool.com/blog/openaxialselective (visited on 02/23/2022).

[21] A. Immonen, E. Ovaska, and T. Paaso. "Towards certified open data in digital service ecosystems." In: *Software Quality Journal* 26.4 (2018), pp. 1257–1297. ISSN: 15731367. DOI: 10.1007/s11219-017-9378-2.

[22] Anne Immonen, Marko Palviainen, and Eila Ovaska. "Requirements of an Open Data Based Business Ecosystem". In: *IEEE Access* 2 (2014), pp. 88–103. DOI: 10.1109/ACCESS.2014.2302872.

[23] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. "Data Scientists in Software Teams: State of the Art and Challenges". In: *IEEE Transactions on Software Engineering* 44.11 (2018), pp. 1024–1038. DOI: 10.1109/TSE.2017.2754374.

[24] Barbara A. Kitchenham and Shari L. Pfleeger. "Personal Opinion Surveys". In: *Guide to Advanced Empirical Software Engineering*. Ed. by Forrest Shull, Janice Singer, and Dag I. K. Sjøberg. London: Springer London, 2008, pp. 63–92. ISBN: 978-1-84800-044-5. DOI: 10.1007/978-1-84800-044-5_3.

[25] Clément Labadie, Christine Legner, Markus Eurich, and Martin Fadler. "FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs". In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*. Vol. 1. 2020, pp. 201–210. DOI: 10.1109/CBI49978.2020.00029.

[26] Johan Linåker and Per Runeson. "How to Enable Collaboration in Open Government Data Ecosystems: A Public Platform Provider's Perspective". In: *JeDEM - eJournal of eDemocracy and Open Government* 13.1 (Aug. 2021), pp. 1–30. DOI: 10.29379/jedem.v13i1.634.

[27] Dominik Lis and Boris Otto. "Data Governance in Data Ecosystems – Insights from Organizations". In: July 2020. URL: https://aisel.aisnet.org/amcis2020/strategic_uses_it/strategic_uses_it/12/.

[28] Bardi Matturdi, Xianwei Zhou, Shuai Li, and Fuhong Lin. "Big Data security and privacy: A review." In: *China Communications, Communications, China, China Commun* 11.14 (2014), pp. 135–145. ISSN: 1673-5447. DOI: 10.1109/CC.2014.7085614.

[29] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. "Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process". In: *Organization* 1.2 (2022), p. 3. DOI: 10.48550/arXiv.2110.10234.

[30] Marcelo Iury S. Oliveira, Glória de Fátima Barros Lima, and Bernadette Farias Lóscio. "Investigations into Data Ecosystems: a systematic mapping study." In: *Knowledge & Information Systems* 61.2 (2019), pp. 589–630. ISSN: 02191377. DOI: 10.1007/s10115-018-1323-6.

[31] Marcelo Iury S. Oliveira and Bernadette Farias Lóscio. "What is a Data Ecosystem?" In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. dg.o '18. Delft, The Netherlands: Association for Computing Machinery, 2018. ISBN: 9781450365260. DOI: 10.1145/3209281.3209335.

[32] Per Runeson. "Open Collaborative Data - using OSS Principles to Share Data in SW Engineering". In: *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. 2019, pp. 25–28. DOI: 10.1109/ICSE-NIER.2019.00015.

[33] Per Runeson, Emelie Engström, and Margaret-Anne Storey. "The Design Science Paradigm as a Frame for Empirical Software Engineering". In: *Contemporary Empirical Methods in Software Engineering*. Ed. by Michael Felderer and Guilherme Horta Travassos. Cham: Springer International Publishing, 2020, pp. 127–147. ISBN: 978-3-030-32489-6. DOI: 10.1007/978-3-030-32489-6_5.

[34] Per Runeson and Martin Höst. "Guidelines for conducting and reporting case study research in software engineering." In: *Empirical Software Engineering* 14.2 (2009), pp. 131–164. ISSN: 1573-7616. DOI: 10.1007/s10664-008-9102-8.

[35] Per Runeson, Thomas Olsson, and Johan Linåker. "Open Data Ecosystems - an empirical investigation into an emerging industry collaboration concept." In: *Journal of Systems and Software* (2021). ISSN: 0164-1212. DOI: 10.1016/j.jss.2021.111088.

[36] *Secoda - Data discovery built for modern data teams*. URL: https://www.secoda.co/ (visited on 01/03/2022).

[37] Syed Iftikhar Hussain Shah, Vassilios Peristeras, and Ioannis Magnisalis. "Government Big Data Ecosystem: Definitions, Types of Data, Actors, and Roles and the Impact in Public Administrations". In: *J. Data and Information Quality* 13.2 (May 2021). ISSN: 1936-1955. DOI: 10.1145/3425709.

[38] David V. Thiel. *Research Methods for Engineers. [Elektronisk resurs]*. Cambridge University Press, 2014. ISBN: 1107034884.

[39] D.W. Turner III. "Qualitative interview design: A practical guide for novice investigators." In: *Qualitative Report* 15.3 (2010), pp. 754–760. ISSN: 10520147.

[40] Allison Viola and Shefali Mookencherry. "Metadata and Meaningful Use". In: *Journal of AHIMA* 83.2 (2012), pp. 32–38. ISSN: 10605487.

[41] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 15, 2016). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Publication characteristics;Research data Subject_term_id: publication-characteristics;research-data, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[42] A. Yoon. "Red flags in data: Learning from failed data reuse experiences." In: *Proceedings of the Association for Information Science and Technology* 53.1 (2016), pp. 1–6. ISSN: 23739231. DOI: 10.1002/pra2.2016.14505301126.

[43] Ehtisham Zaidi, Guido De Simoni, Roxane Edjlali, and Alan D Duncan. "Data Catalogs Are the New Black in Data Management and Analytics". In: *Gartner, Consultancy Report* (2017).

[44] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. "Innovation with open data: Essential elements of open data ecosystems." In: *Information Polity: The International Journal of Government & Democracy in the Information Age* 19.1/2 (2014), pp. 17–33. ISSN: 15701255. DOI: 10.3233/IP-140329.

[45] Anneke Zuiderwijk, Marijn Janssen, Jing Zhang, Gabriel Puron-Cid, and J. Ramon Gil-Garcia. "Towards decision support for disclosing data: Closed or open data?." In: *Information Polity: The International Journal of Government & Democracy in the Information Age* 20.2/3 (2015), pp. 103–117. ISSN: 15701255. DOI: 10.3233/IP-150358.

Appendices

Appendix A

Literature Study - Keywords

Conceptual areas of interest:

- A1. Value
- A2. Intrinsics
- A3. Governance
- A4. Evolution

General and area-related keyword-string:

- General: (Challenge* OR Issue*) AND (Data Ecosystem OR Data Catalog* OR Data Management OR Data Collaboration OR Data Collaborate OR Data Discover*)
- A1: (Challenge* OR Issue*) AND (Data Value OR Data Sharing OR Data Reuse OR Metadata Value OR Metadata Sharing OR Metadata Reuse)
- A2: (Challenge* OR Issue*) AND (Data Ecosystem Taxonomy OR Data Ecosystem Intrinsics OR Data Ecosystem Conceptualization OR Data Ecosystem Interoperability)
- A3: (Challenge* OR Issue*) AND (Data Governance OR Data Governance requirement* OR Data Governance privacy OR Data Governance industry OR Data Governance industrial)
- A4: (Challenge* OR Issue*) AND (Data Ecosystem OR Data Ecosystem evolution OR Data Ecosystem tool* OR Data Ecosystem industry OR Data Ecosystem industrial)

Appendix B

Interview - Questions

* - Specifically for Data Producers and Provider
** - Specifically for Data User and Re-user
*** - Specifically for Legal and Keystone Actors

Part 1 - Background

- What is your role at Axis?
- How do you work with data as of today?
  - Can you please describe a bit about your work in regards to data?

Part 2 - Value

- According to you, from your point of view, what is the value of the data you work with? What is it used for?
- Is collaboration on data between departments active today?
- Are there any limitations connected to working with the data as of today?

Part 3 - Intrinsics

- Do you often add new types of data to your collection?*
- Do you produce metadata around the data?*
- To which extent do you find data by the use of metadata?**
- Are there data quality and quantity requirements documented?**
- What would be important attributes of the data when accessing its usefulness?
- Do you produce metadata around the data?*

- To which extent do you find data by the use of metadata?**
- Are there any legality concerns connected to the data you work with?
  - How do you know about legal concerns connected to the data?
- What would be important attributes of the data when accessing if it is good to use or not to use?***

Part 4 - Governance

- Have you requested data from any other department or individual than your own?**
- Have you been asked for data from another department?**
- How easy is it to find and evaluate data sources as of today?**
- Are you aware of what data exists in other departments?**&***
- Do you share your data with other departments?*
- Who owns the data?*&***

Part 5 - Evolution

- How well is the department and organizational standardization of data collected from a legal perspective?***
- Do you use any technical platforms today for accessing/finding data?**
- What do you see as the potential for a technical platform working and collaborating on data in the future?
  - Increase the value of the data?
  - Increase discoverability?
  - Sharing your data?*
  - Finding and comparing the best source to use?**
  - Handle legality?***

Part 6 - End

- Do you have anything that you think is important to say regarding this subject?

Appendix C

Interview - Coding

C.1 Value

Value of Data					
Knowledge	P (7/9)	Improvements	P (6/9)	Product	P (6/9)
Advise	6	Future	2, 4	Product	4, 8, 9
Knowledge	2, 4, 8, 9	Quality	4, 6	Service	3
Understanding	4	Improvements	2, 5, 6, 8	Data Users	2
Insights	8	Describing	3	Report	7
Analysis	7	Stability	3		
Overview	1, 6, 7	Changes	5		
Features	P (2/9)	Limitation	P (5/9)	Future Value	P (3/9)
Features	2	Resources	9	Later Value	3, 8
Behaviour	4	Quantity	4, 8	Promoting	4
		Data	5, 7		

Table C.1: Coding and grouping of *Value of data*.

Value of Collaboration					
Limitations	P (5/9)	Mature	P (3/9)	Improvements	P (3/9)
Quantity	7	Common ground	8	Sharing	5
Maturity	2, 4	Collaboration	8	Optimization	8
Difference	8	Maturing	4	Effectiveness	4
Data	7	Overview/insight	3	Improvements	5
Knowledge	3				
Silos	7, 8				
Terminology	7				
Variety	7				
Future Value	P (5/9)	Stability	P (2/9)	Integration	P (1/9)
Increased Value	2, 8	Monitoring	6	Integration	4
Future Value	4, 5, 8	Releases	5	Combine	4
Promoting	6				

Table C.2: Coding and grouping of *Value of Collaboration*.

C.2 Intrinsics

Type of data					
Characteristics	P (5/9)	Dependency	P (4/9)	Feature	P (1/9)
Metadata	8	Use-cases	8	Automatic	5
Personal	6	Origin	5		
Logs	1	Quantity	5, 7		
Telemetry	3	Variation	5, 7		
XML	7	Frequency	3		

Table C.3: Coding and grouping of *Type of data*.

Quality					
Description	P (5/9)	Relationship	P (5/9)	Analyse	P (3/9)
Schema	1	Feedback	7	Analyse	4
Description	8	Trust	4	Manual	9
Definition	2, 3	Users	2	Usage	3
Metadata	9	Use-cases	3, 8	Metrics	4
Knowledge	P (3/9)	Filter	P (2/9)	Completeness	P (5/9)
Maturity	4	Correctness	8	Quantity	6
Planning	8	Cleaning	2, 4	Completeness	2, 4, 5, 8
Knowledge	4, 5				

Limitation	P(7/9)
Conditions	8
Curentness	3, 9
Responsibility	1
Cost	6
Changes	5
Knowledge	7
Human	7

Table C.4: Coding and grouping of *Quality*.

Legal					
Access Control	P (4/9)	Rules	P (4/9)	Limitation	P (6/9)
Control	4	Licenses	8	GDPR	1
Access Control	7	Contract	4	Linked data	2, 6
Use-cases	4	Instruction	6	PII	6
Approval	1	Transparency	8	Grey Zone	5
Review	3	Responsibility	4, 6	Uncertainty	4
Region	7	Maintain	5	Sensitive	8
Abort	3				
Characteristics	P (5/9)	Anonymization	P (2/9)	Knowledge	P (3/9)
PII	5, 7	Aggregation	7, 8	Advice	2, 9
Metadata	8	Analysis	8	Knowledge	5
Linked data	5	Cleaning	8		
Anonymous	4, 5, 6				

Table C.5: Coding and grouping of *Legal*.

C.3 Governance

Acquisition					
Processes	P (6/9)	Responsibility	P (4/9)	Limitation	P (4/9)
Process	3	PII	7	Legacy	2
Collaboration	8	Responsibility	5, 6	Names	5
Verbal	7	Legal	6, 7	Update	5
Tool/Platform	5	Cleaning	7	Discoverability	2, 7, 9
Meeting	9	Schema	1	Tool specific	2
Communication	4				
Access Control				P (3/9)	
Request Access				3	
Access Requirements				2	
Use Case				8	

Table C.6: Coding and grouping of *Acquisition*.

Relationships					
Communication	P (4/9)	Roles	P (5/9)	Future Value	P (2/9)
Meetings	2, 4	Roles	1, 2, 7	Need	1
Communication	2, 5	Responsibility	3, 8	Linked data	5
Promoting	5	Access Control	1, 8		
Platform	9				
Limitation				P(5/9)	
Feedback				3, 9	
Knowledge				3, 7	
Discoverability				7	
Access control				7	
Communication				4	
Discussion				7	
Overview				4, 7	
Maturing				8	
Legal				8	

Table C.7: Coding and grouping of *Relationships*.

Competition					
Knowledge	P (2/9)	Limitation	P (5/9)	Product	P (2/9)
Documentation	5	Knowledge	1, 2, 4	Service	5
Communication	3, 5	Silos	7	Evince	4
Knowledge	5	Discoverability	1, 2, 5		

Table C.8: Coding and grouping of *Competition*.

C.4 Evolution

Maturity					
Terminology	P (3/9)	Limitation	P (4/9)	Features	P (2/9)
Common	8	Discoverability	3	Find Duplications	3
Centralisation	8	Linking data	6	Link data	5
Terminology	7	Legacy	2, 3	Data linage	7
Discussion	2, 7	Different Schema	5	Browse	3, 5
Save Time	2	Feedback	2	Sort	5
Define Roles	7	Linking to PII	6	Searchable	3, 5
Standardisation	8				
Discoverability	P (5/9)	Knowledge	P (7/9)	Monitoring	P (4/9)
Discoverability	1, 2, 3, 7	Documentation	5, 7	Use case	4, 6, 7
Metadata	7	Mature	8	Usage	2
Transparency	1, 2, 7	Improvements	7	Purpose	6
Trust	5, 7	Knowledge	1, 2, 3, 4, 5, 7	Feedback	2
Open	7	Value Generation	8	Objective	4
Access Control	P (5/9)	Platform	P (3/9)	Overview	P (5/9)
Region	7	Collaboration	7	Overview	1, 2, 3, 7
Aggregation	8	Platform	9	Communication	7
Handle Legal	5, 6	Service	8	Perspective	5
Access Control	3, 7				

Table C.9: Coding and grouping of *Maturity*.

Appendix D

Validation Questionnaire

Master's Thesis at DDM - Data Ecosystem, follow-up from interview

We have now done our literature study, case study (interviews) and finally a mapping of the challenges found, as requirements, toward a technical platform for supporting a Private* Data Ecosystem**.

As validation we would like for you to answer this from your perspective; how relevant you think each of the 14 requirement are.

1, Strongly disagree
2, Disagree
3, Neutral
4, Agree
5, Strongly agree

The results of this survey will be anonymously presented in the final report.

* Private in the meaning of being within the company (Axis).
** A Data Ecosystem can be defined as: "socio-technical complex network in which actors interact and collaborate with each other to find, archive, publish, consume, or reuse data as well as to foster innovation, create value, and support new businesses".

* Required

1. The platform shall support and present a standardized environment around its components.
*

the standardization of the platform around terminology, e.g., roles, relationships, data intrinsics and actions may help the actors to conform to a coherent understanding of the landscape.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

2. The platform shall support transparency around who uses the data and for what.
*

transparency may support insights around how the data is used and presented, further providing feedback between the involved parties.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

3. The platform shall support a standardized exchange reference model for different types of data.
*

by supporting a common exchange reference model applicable for different types of data, support for interoperability between different data types may increase.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. The platform shall support the extraction of meta-data from the dataset.
*

as a requirement to allow for the assessing the dataset anonymously, further allowing for contracts if data is wanted.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. The platform shall support the presentation of meta-data around the dataset.
*

as a requirement to allow for the assessing the dataset anonymously, further allowing for contracts if data is wanted.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. The platform shall support monitoring of the evolution of the data.
*

where the evolution of data includes, but is not limited to: The usage, classification, linage and, transformation as updates and changes of the dataset.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. The platform shall support communication around the usage of data.
*

the platform should include communication around the data in order to assess the quality in relation to its context of usage.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. The platform shall support statistical quality checks of the data.
*

in order to support the current way of working, the platform should address the existence of quality checks as missing values or outliers.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. The platform shall support the linkage of documents between actors and related to the data.
*

standardized quality agreement around the data quality, i.e., such as SLA's.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

10. The platform shall support the presentation of documentation around the data.
*

for the acquisition of data, information around licenses, contracts, and other processes and information of whom to contact should be provided to establish knowledge of the data and acquisition process.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

11. The platform shall support the discovery of existing resources.
*

data sources within the ecosystem should be searchable, findable. and request access is an essential part of a data ecosystem and should be handled by the technical platform.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

12. The platform shall base its governing mechanisms on established, common roles.
*

if new governance mechanisms, e.g., regarding access rights and responsibilities, are introduced or transferred onto the platform, these should be based on the ecosystem's common roles due to the importance of a common understanding.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

13. The platform shall support feedback and discussion around the usage of the data sources.
*

as communication and feedback are important parts of relationships, the platform should support the possibility for feedback and discussion around the data.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

14. The platform shall support transparency around the provision and usage of data.
*

as communication and feedback are important parts of relationships, the platform should support the possibility for feedback and discussion around the data.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

15. Any important factors/requirements you think we missed? If yes, please explain.

16. Other comments

17. I would like a copy of the final master's thesis sent to me once done. *

- Yes
- No

Appendix E

Questionnaire - Results

	P1	P2	P3	P4	P5	P7	P8	P9	*Average Score*
Req. 1	4	5	5	4	5	4	3	5	*4,38*
Req. 2	5	5	5	5	4	4	2	5	*4,38*
Req. 3	4	5	4	4	4	4	5	4	*4,25*
Req. 4	5	5	4	5	4	3	5	3	*4,25*
Req. 5	5	5	4	3	4	3	5	3	*4*
Req. 6	4	5	5	3	4	4	5	4	*4,25*
Req. 7	5	5	4	3	4	3	5	3	*4*
Req. 8	5	4	5	5	4	3	3	5	*4,25*
Req. 9	4	3	5	3	3	3	2	4	*3,38*
Req. 10	5	5	5	3	4	4	3	4	*4,13*
Req. 11	4	5	5	3	3	4	5	5	*4,25*
Req. 12	5	5	4	2	4	3	3	4	*3,75*
Req. 13	5	5	4	2	4	3	4	4	*3,88*
Req. 14	5	5	4	4	4	4	5	4	*4,38*
Average Score	*4,64*	*4,79*	*4,5*	*3,5*	*3,93*	*3,5*	*3,93*	*4,07*	***4,11***

Table E.1: Questionnaire results from all the participants.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2022-02-17

EXAMENSARBETE Data Ecosystem as a solution for Intra-Organizational Data Sharing: Benefits, Challenges and Requirements
STUDENTER Christian Bilevits, Adam Hägglund
HANDLEDARE Per Runeson (LTH), Johan Linåker (LTH), Anton Friberg (Axis Communications)
EXAMINATOR Martin Höst (LTH)

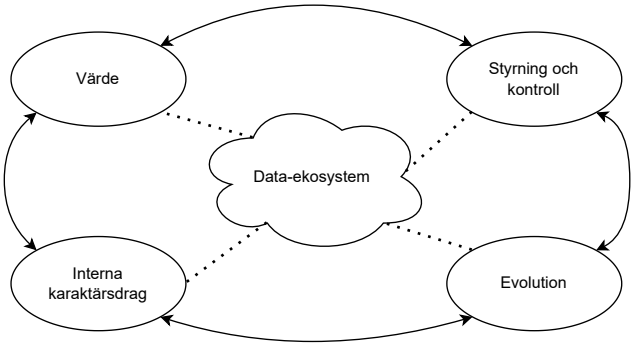
Är data-ekosystem vägen framåt för datadelning inom företag?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Christian Bilevits, Adam Hägglund**

Data blir allt viktigare för företag och samhälle, där alltmer data samlas in. Men hur kan den delas så att den kommer till nytta? Detta arbete har utforskat fördelar och utmaningar med datadelning inom ett företag, samt tagit fram plattformskrav.

Data har blivit en alltmer viktig och värdefull resurs som både återfinns och eftersöks i många delar av vårt samhälle, inte minst inom den tekniska industrin. Data har visat sig kunna bidra till ny innovation genom exempelvis analys, maskininlärning och artificiell intelligens vilket även har lett till ett ökat behov. Trots detta så betyder det inte att data är lätt att få tag på eller att dela med sig av, då delning av data innefattar både praktiska, legala och tekniska svårigheter. Mot bakgrund av dessa problem så har ett relativt nytt koncept börjat få mer fokus i forskningsvärlden, ett koncept under namnet data-ekosystem. Ett data-ekosystem beskriver ett socio-tekniskt kollaborativt nätverk av aktörer där alla är intresserade av just att dela och samarbeta kring data.

I vår forskning tog vi en tolkning av detta koncept och utforskade den befintliga litteraturen samt ett företag där intresset, men även svårigheter kring datadelning har upplevts. Företaget i fråga är ägare till mycket data och ser en potential i att dela sin data mellan sina olika avdelningar. Målet med vår studie var att kartlägga vilka fördelar och svårigheter som kommer med datadelning inom ett data-ekosystem, och utifrån våra fynd ta fram en kravbild att ställa mot en teknisk plattform för att stödja ett sådant ekosystem för datadelning inom företaget.



Figur 1 – Modell över det data-ekosystem koncept som utforskades.

Vårt resultat visade att det fanns många likheter både när det gäller fördelar och utmaningar mellan företaget och den befintliga litteraturen. Ett data-ekosystem visade sig tydliggöra värdet av samarbete, där en gemensam miljö med standarder är nödvändig för att främja värdeskapande, kunskap, transparens samt även hantering av kvalitet och legalitet kring data. Något som framgick att vara av extra vikt hos företaget var även kommunikation i flera olika former, i positiv och negativ bemärkelse. Utifrån resultaten tog vi fram 14 olika krav som kan ställas på en teknisk plattform för ett data-ekosystem, dessa utvärderades mot befintliga plattformar vilket visade sig lovande.