Inter-Organizational Data Sharing Processes - An Exploratory Analysis of Incentives and Challenges

Malysh, Konstantin; Ahmed, Tanvir; Linåker, Johan; Runeson, Per

Published in:
Proceedings - 2024 50th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2024

DOI:
10.1109/SEAA64295.2024.00021

2024

Inter-organizational Data Sharing Processes – an exploratory analysis of incentives and challenges

Konstantin Malysh
Dept. of Computer Science
Lund University
Lund, Sweden
0000-0002-3659-3093

Tanvir Ahmed
Dept. of Mgmt and Engineering
Linköping University
Linköping, Sweden
0000-0002-4798-4823

Johan Linåker
Research Inst. of Sweden
RISE
Lund, Sweden
0000-0001-9851-1404

Per Runeson
Dept. of Computer Science
Lund University
Lund, Sweden
0000-0003-2795-4851

***Abstract*—Businesses across different areas of interest are increasingly depending on data, particularly for machine learning (ML) applications. To ensure data provisioning, inter-organizational data sharing is proposed, e.g. in the form of data ecosystems. The aim of this study was to perform an exploratory investigation into the data sharing practices that exist in business-to-business (B2B) and business-to-customers (B2C) relations, in order to shape a knowledge foundation for future research. We launched a qualitative survey, using interviews as data collection method. We conducted and analyzed eleven interviews with representatives from seven different companies across several industries with the aim of finding key practices, differences and similarities between approaches, so we could formulate the future research goals and questions. We grouped the core findings of this study into three categories: organizational aspects of data sharing, where we noticed the importance of data sharing and data ownership as business driver; technical aspects of data sharing, related to data types, formats, maintenance and infrastructures; and challenges, with privacy being the highest concern along with the data volumes and cost of data.**

***Index Terms*—Data sharing, machine learning, data engineering, B2B and B2C practices, empirical interview study.**

I. Introduction

The trend towards data-driven business is prevalent across private and public organizations. The transition towards data-driven business is typically connected to a transition towards servitization, i.e., a process where companies transfer from offering distinct products, to bundling products with data-driven services. Tronvoll et al. conclude that *"new data-sets enable novel services and ultimately enhance competitive advantage. However, issues related to the data generation, collection, utilization, and ownership may create new tensions between firms."* [1]

Inter-organizational data sharing between businesses is therefore proposed as a potential mitigation to ensure relevance and efficiency in data supply related to their business goals. For example, Coyle et al. argue in their Value of Data report that, *"[v]alue comes from data being brought together, and that requires organizations to let others use the data they hold"* [2]. This does not imply that all data becomes publicly open. They define a Data Spectrum, ranging from Closed via Shared to Open, which can be balanced depending on the incentives for sharing and governed through relevant licensing and management procedures and tools.

To better understand incentives and challenges for companies to share data between them (business-to-business, B2B) or with their customers (business-to-customers, B2C), we launched an exploratory qualitative survey, using interviews for data collection. This exploratory study is our first step to address the following global research question: *What is the current state (practices, experiences, challenges) of data sharing as the driver of the informational flow between companies?* Our interests embrace both business and technical data sharing aspects.

We have defined data sharing as a process of exchange of data between *receiver(s)* and *sender(s)*, no matter whether they are a business or an individual. Any process that has been established in order to support the data sharing process is defined as a *data sharing practice*; this definition includes, for example, laws and regulation on data, development of sufficient infrastructure, developing methodology and algorithms, etc. If an organization uses data sharing practices and considers it as a driver for their business, such company is considered *data-driven*.

We conducted interviews with employees of seven companies across different domains (see Figure 3). We found that while all the companies value data sharing as a driver of their business, there are similarities and differences between their approaches on practices that help to use the data effectively and various common challenges that they have met. Based on the findings, we drew conclusions on what practices are more diverse and need more thorough investigation in order to help the said companies organize their processes in a more efficient way. The key points being discussed were the privacy concerns, collaboration relationships, assessing data and data sharing using metrics and overall data handling withing the organizational ecosystems.

II. Related work

To provide an overview of existing research, Oliveira et al. [3] surveyed the literature on data sharing between organizations from an *ecosystem perspective*. They highlight how these organizations can act as orchestrators, intermediaries, and/or provide a common platform and marketplace for the data sharing, and related artifacts. Runeson et al. [4] focus specifically on data ecosystems where the data is shared openly

2376-9521/24/$31.00 ©2024 IEEE
DOI 10.1109/SEAA64295.2024.00021

under public licenses, highlighting the need for intrinsics, and a governance structure that creates trust among the ecosystem actors. Linåker and Runeson [5] further point to the need of a central and neutral actor facilitating the collaboration, and also enabling the data sharing and adoption through common infrastructure and processes. Elgarah et al. [6] propose different views on the data exchange in inter-organizational relationships, connecting the data-related processes with market governance, trading relationships, integration, globalization, and state the importance of managers understanding the trends and their readiness to react. We are trying to expand the scope of knowledge by receiving more practical information on real-world data sharing interactions.

Challenges related to the data sharing processes are an important topic and are widely presented in literature. Olsson and Bosch [7] explore challenges experienced by embedded software companies when moving from traditional to continuous data collection and management practices, and adopting service-oriented business models. Combining qualitative and quantitative data, internal and external sources, and defining meaningful metrics are three of the more prominent challenges. Aaltonen et al. [8] discuss the data sharing processes from a range of different standpoints, such as the importance of understanding the data governance, significance of focusing on demand side along with the supply side of data-related research, need of a social science of data in order to investigate data as a center of social settings. In our exploratory research, we have the goal to find connections and patterns between different types of challenges, in order to see how those could be solved from both technical and business points of view.

Some of the related publications take a *data scientist*'s perspective approach. Kim et al. [9] investigate challenges internally, highlighting issues related to the limited availability of data, and the low quality of the data collected. Munappy et al. [10] look at the data management process for deep learning models, also pointing to limited availability of labeled and high quality data, and of tools and processes for sharing and processing of the data. One of the goals of our study is to unite the technical and business point of views to help those two worlds be more productive in achieving goals through the means of data sharing.

Several *single case studies* are also present: for example, Hüner et al. [11] investigate a case of data sharing in Beiersdorf, talking about data defects and defining metrics for product data quality measurement. Gelhaar et al. [12] explore the motives and incentives of data sharing in industrial data ecosystems through the model of motives and incentives being behavior activators. Importance of every ecosystem participant sharing and value of individual data sharing use cases are being underlined. Brechtel et al. [13] also discuss the topic of data sharing for industrial data ecosystems from a socio-technical perspective, defining the core challenges, such as data availability, lack of knowledge, poor venture-driven mindset, unsuitable incentive system, data responsibilities, poor accessibility of use cases and high initial investments. In our research, we worked on finding an interplay between different areas of industry to enable data sharing across the boundaries of the fields.

From the *business point of view*, D'Hauwers et al. [14] among others define and compare data sharing business models frameworks for intra- and inter-organizational data sharing, and state that the core factors influencing data sharing collaborations are value creation, data governance, as well as ecosystem trust and data trust. In this paper, we are trying to assess the business-oriented means related to data sharing across different areas of industry, and are also figuring out interactions between the business goals and technology side of the project by approaching industry representatives from either side and discussing the cooperations between two approaches.

III. Research design

The research presented in this paper is an exploratory qualitative survey [15], based on interviews with eleven representatives of seven different companies with the goal of investigating the data-related practices. Despite the variety of domains and interests of these organizations, they are all data-driven and are exploring data sharing practices suited for their business. As one of the goals of this study is to find some patterns in the data-related behaviour of companies, an interview study approach has been selected due to its suitability for such task [16]. Our research process is inspired by the process steps proposed by Strandberg [17], with Figure 1 representing the outline of this study.

A. Interviews

Eleven persons have been contacted for the participation in the study, each representing one out of seven organizations in different business domains. We have purposely selected companies and persons from our industry network, based on their interest and experience in data-driven businesses. Even though all interviewees are based in Sweden, the corporations are all international. Table I contains brief information about the interviewees; the enumeration is done in chronological order of the interviews.

B. Data collection

Eleven semi-structured interviews have been conducted with the selected representatives, during April–May 2023 and March-April 2024. The interviews have lasted 45 to 90 minutes, and were fully online, except I6 which was hybrid (interviewees and Authors 1 and 4 in one room, Author 2 remote). Authors 1 and 2 have participated in all interviews and have been the main drivers of the technical and business parts in the interview, respectively. Authors 3 and 4 have participated interchangeably.

The interviews followed an interview guide[1] covering interviewee's experience and data sharing practices they use, although loosely followed as long as interviewees provided relevant information for the study. The interviews have been recorded and stored locally for the further steps of processing.

[1]https://figshare.com/s/acc4528aaa33ea5b5c5d

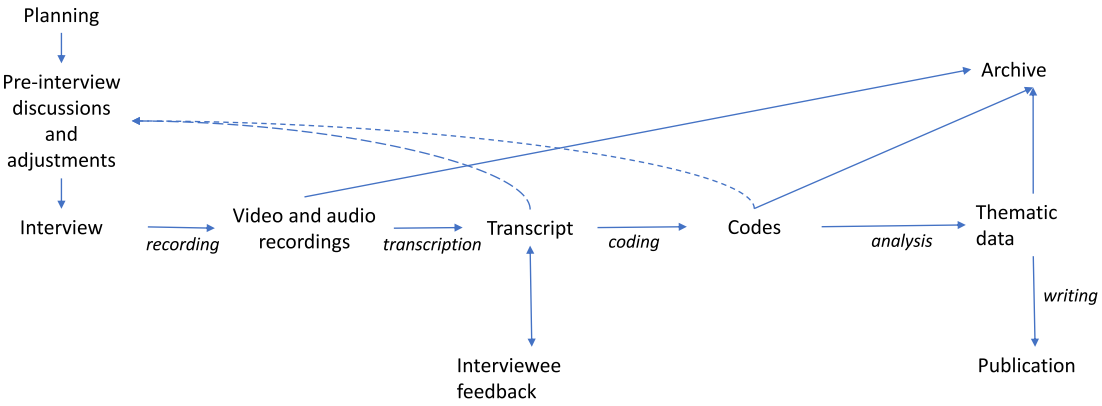

Fig. 1. Graphical representation of the different iterations of the study, inspired by Strandberg [17].

TABLE I
INFORMATION ABOUT THE INTERVIEWEES.

Company ID	Interviewee ID	Interviewee area of interest	Company domain
C1	I1	Data Engineer	Video surveillance
	I7	Engineer Manager	
C2	I2	Physicist and Data Analyst	Multi-disciplinary research facility
C3	I3	Data Engineer	Outdoor power products
	I8	Team Leader in Big Data	
C4	I4	Product Manager and Planner	Heavy Transport
C5	I5	Product Owner	Networking and telecommunications
	I9	Product Owner	
C6	I6	Co-founder and CEO	Advanced robotics
C7	I10	AI Lead	Driving management solutions for vehicles
	I11	IP and Legal Officer	

C. Data analysis

For data analysis, the interviews have been transcribed using a speech-to-text tool, and then sent back to the interviewees for validation. Then the transcripts were coded separately by Authors 1 and 2 using a descriptive coding approach: the core topics and the corresponding quotes by the interviewees have been highlighted and transformed into the corresponding codes [18]. After the initial coding process of I1–I6, a synchronisation meeting between the Authors 1 and 2 has taken place, with the help and mentorship and coding validation by the Author 3, including agreeing on the sufficient saturation of the codes. As a validation set, I7–I11 were then coded using the same scheme, and found being sufficient. The resulting codes have been analyzed and the results of the analysis are presented in the next chapter.

IV. FINDINGS

During the process of the interviews, we have identified that the companies whose representatives have agreed to take part in the research, are indeed heavily data-driven and have developed data sharing practices over time, whether those are infrastructures, algorithms or regulations. Due to the rapid technology development, their practices are also bound to change, implying the increase in the maturity and knowledge about data sharing.

We have split the findings into three core subcategories: i) the topics related to the *organizational* part of the process; ii) the topics that are directly *linked to the data and the technical side of data sharing*; and iii) *overall challenges* linked to the process as a whole. Figure 3 maps the findings of the study and relates them to each interview.

We identified two rather distinct types of data, namely i) the core customer or application data, and ii) the operations or system monitoring data. In the first category, we find video streams in the surveillance domain, vehicle positions in automotive, research experiment data in the research facility etc. In the second category, we find computational statistics, system and communication load, control system signals etc. These data types are discussed in this section.

A. Organizational data sharing aspects

1) Collaboration aspects: **Business drivers:** Collaborations with other businesses or individual customers are the core of the data sharing processes in the scope of this study. All the interviewees have confirmed the importance of collaborating with other entities in order to increase productivity and to gain benefits, whether it is financial profit, technology boost or higher connection to the academia.

The study participants have mentioned that data sharing is a communication and business enabler, either through their services and workshops, where the customers could test or

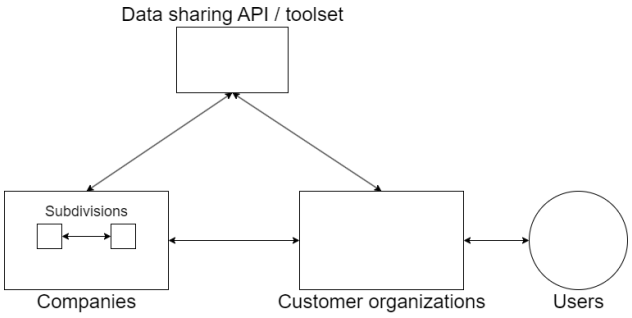

Fig. 2. Conceptual model of organizational entities involved in data sharing processes

perform the product diagnostics; or just as a business communication tool, setting a connection between several parties on their common interests. **I3** has said that *"We want to make use of data, act more data-driven than we do today. ... We're looking forward to more data, even though it brings costs, we still see the value, and we don't expect it to decrease, rather the opposite"*, while **I2** has pointed out in relation to their connection with their peers: *"We could have more peers if we could share the data a little bit easier."*. **I6** has stated the importance of conscience of collaborations in relationship to the data collection itself: *"We want to avoid ending up in these situations where you get something and you can't really analyze it and use it in a proper way except for that purpose, specific purpose."*

Technology drivers: The technological progress is considered to be one of the key drivers for the better data sharing techniques, with industrialization and shift to cloud storage as reasons for the higher data sharing standards, as confirmed by several interviewees. For example, **I5** states that the *"Things are getting more and more cloudified, ..., cloud is giving more opportunities to actually do observations and control the system compared to if you run on a dedicated box for compute"*, while **I1** has mentioned that in general they are shifting towards more data sharing after learning about the data-related processes and regulations.

Relationships: Another interesting observation is that all the seven companies follow the same inter-organizational model of data sharing: there are two large hubs, companies themselves and their customers; there are data sharing processes inside the companies and between companies and their customers, directly or via some framework; and then there is also another data sharing process, between the customers and individual users, as, for example, with workshops enabling diagnostics and maintenance for the users for **I3**, or academia and students through the academic projects for **I2**. The concept of the model is presented in Figure 2, which emerged from the interviewees' descriptions of actors involved in their data collection and sharing processes.

The arrows in the model represent the data flow; however, some of the challenges stated later in this paper are also flowing along the arrows: privacy, geopolitics and cost-related challenges travelling mostly along the external arrows and data volumes being represented everywhere.

2) Metrics for the quality assessment: Opinions on how to measure the quality of data sharing from the business point of view are different. Some internal key performance indicators (KPIs) have been mentioned without much detail, however, the fact of presence of such KPIs already indicates the importance of high-quality data sharing. **I1** has mentioned feedback assessment as their core metric due to the importance of the B2C relationships: *"For me, it's quite simple, but you might not like to hear this, but it's listening to your customers. So for me, basically I try to talk with the people who actually use this data, the people who are extracting value from this data."*, and **I2** has mentioned just having some non-specific metadata to evaluate the quality of the data they have, as metadata makes data more understandable. **I8** has also outlined the amount of shares of specific data on a monthly basis.

3) Data ownership: The topic of who owns what data is important nowadays, especially with new regulations gradually coming in force, and the interviewees have confirmed this. In general it has been mentioned that the ownership question is lighter inside the companies due to lack of legal challenges from outside. As for the B2B and B2C data sharing processes, one of the approaches mentioned by **I3** is granting the raw original data ownership to the customer and granting the processed data ownership to the organization; however, as specified by **I5**, it can sometime be hard to clearly define the data proxy in their business, for example, with use of end-user agreements.

I6 has also pointed out the importance of differentiating different ownership-related terms while describing work of some of their systems: *"So everyone owns the data they have produced and that should be possible to mark it up like owned. But everyone working in that system should have* access *and do their way of aggregation and learning their type of things. ... It's important to make the distinction, distinction between owning the data and accessing it and using it."* They further stress that it is important to be able to provide different types of data ownership to, for example, the integrators of the products.

Overall, data ownership is considered to be an important topic by the interviewed representatives and is needed to be investigated in more detail, potentially from different data ownership knowledge maturity point of view.

B. Technical data sharing aspects

In this subsection, we present the technology-related side of data sharing processes. During the interviews, the following core patterns have been discussed: i) data collection, as in the initial data sharing process; ii) data maintenance, being processing and analyzing the collected data, and iii) further data exploitation, using the results of the work on the data for the business benefits.

1) Data and platform: **Type of data:** The data itself that is being collected by a company defines the data sharing processes, how and what for the data will be used. All the

interviewees have mentioned a wide variety of data instances they are collecting, such as user logs, sensor data or equipment geolocation, forming a large selection of data to be handled differently.

Infrastructure for sharing: There is a large assortment of technologies being used for the data collection and exchange, usually being some kind of external data collection mainframes of different complexities with an API in order to access data, and some form of a data storage, with cloud solutions (**I3**, **I5**), local intra-organizational data lakes (**I3**) or data mesh (**I8**) being mentioned as the data aggregation tool. **I1** has mentioned an important requirement for such storage: *"The main tool with working with all of this is to make sure that the data is structured and well understood and managed in a database and that every field is documented"*.

Data formats: As for the data formats, the answers have mostly been vague, with snapshots and logs being stored in some format being mentioned. **I6** has pointed out some important qualities of data in relationship to its format: *"So it depends on what you plan to use with the data for, but I wish we had been using more of this standardized data format for storages and more of semantic tagging of data."*. **I2** has mentioned their way of sharing data with their customers being through e-mails or USB-sticks, which entails obvious restrictions on the data to be shared. In conclusion, the interviewees have talked about large variety of technical approaches for data collection and data formatting of different complexity levels, indicating potential challenges related to initiating a new data sharing process.

2) Data maintenance: After the data is collected, the next step is to start processing it. The importance of the data cleaning has been acknowledged by some participants, as data can come 'dirty' or incomplete; also, performing such task manually is very tedious and time-consuming, so there is a need for automation support to the data cleaning process. As for the actions to be performed on the data itself, a large variety of examples have been presented by the interviewees, as the different approaches are needed in order to solve different data-related tasks. Some of the notable mentions are encryption/decryption processes of different kinds for securing the data: **I5** has mentioned that *"Some customers are completely happy that we store the data encrypted. Some people want us to store it in our premises, some people want us to store it in a bunker with armed guard"*; another one that has been brought up is machine learning algorithms by **I2** and **I4**, as they provide a big set of applications and are also considered emerging in the business and technology field, as **I2** has stated: *"My general view is that, things are happening so fast, so I really hope they are in connection with science. I mean AI, machine learning"*. **I8** has mentioned the importance of testing being a part of the data ecosystem as an analysis tool as well as a data sharing enabler: *"I do believe that testing should be a part of the mesh. I think we need to make it simple for people to share data and to analyze data quality, to make it a very simple thing."*.

C. Data sharing challenges

1) Privacy concerns and the related challenges: **GDPR:** All the interviewees have talked about privacy-related issues as one of the core obstacles for the data sharing process, with GDPR (General Data Protection Regulation, a European Union regulation on Information privacy) being mentioned as the core regulation affecting the data sharing processes, as well as being a powerful shift instrument. Another core way of regulating the sensitive data sharing between a business and a customer is having an end-user agreement or contract regulating what data is being collected and in what way it is being used, being mentioned by multiple interviewees, for example, **I1** has stated that *"if we reach out to a neighbor of one customer and say that other customers in your area have these types of cameras, it cannot be used for that purpose and that's something my team is enforcing actively"*.

Embedded customer information: Both **I10** and **I11** discussed the embedded customer information in the driving data as a privacy challenge. **I10** mentioned, *"For instance, if you know the license plate number of a car and you know the position data, then basically, you know who the owner is or where the owner was and things like that. So that's sensitive information."* Both **I10** and **I11** mentioned protecting customer privacy by anonymizing the customer data at the platform level; and even before they are shared with any actors in the ecosystem. Synthesizing data can be one solution here, which is replacing the real data with synthetic data. For example, reproducing a car registration number as a fictitious number, thus removing the trace of this data back to the customer. **I10** additionally mentioned a more advanced technique for anonymizing data called federated learning for protecting customer privacy while data sharing, entailing using the data to train the platform at the source regarding operational incidents and thus using customer data locally to learn about various driving incidents and send back only the important insights from those learning back to the platform.

Privacy maturity: I3 has also raised a valid point of importance of privacy maturity on the user end: *"One challenge there is, since we're working with thousands of partners and dealers around the globe, and many of these are pretty small companies, like family owned businesses, their awareness of what data sharing means in terms of privacy, security, legal measures is not super high. So we don't know how they treat their network security..."*

Additionally, the data security or governance policies do not facilitate safer data sharing opportunities. **I7** mentioned the lack of technical and legal expertise during these policy formulations as the major barrier to more favorable data governance policies. As mitigation, **I7** pointed out, *"By working very closely with legal, the technology departments can help develop an ethical framework around the data from the beginning. This can create a little bit of safety for working with data."*

2) Data volumes: Throughout the interview process, several interviewees have mentioned the topic of how much

data is being shared and stored in a different light. Mostly the agreement is being reached that the rapidly increasing data volumes are a driver of the innovation, with several interviewees mentioning that always growing data sizes lead to the storage-related challenges; however, **I4** has stated the exact opposite, as the data volumes are still relatively small for their company: *"I would say the data volumes aren't that big still. ... I don't think data speed is that crucial. Actually, I'm not really sure if the technology from that perspective is limiting at the moment. I think it's rather on the application side that someone needs to do the algorithms to understand the data."*.

On a contrasting note, **I10** mentioned that in their company's case, the volume of data is not a challenge. They rather need to decide which data to keep, saying *"It is a very challenging situation, which data to keep and which ones to throw away, because basically not all the data is worth keeping based on their usage purposes."*

3) Geopolitical restrictions: A majority of the organizations' representatives we have interviewed have mentioned their business being global in one way or another, whether it is a direct presence or an international partnership, e.g. through user/customer workshops. **I5** has mentioned the different roles of geopolitics in their data-driven business: for example, the different regional regulations (especially the privacy-related ones) are seen as an obstacle for the business globalization processes. **I5** has mentioned a formidable distinction between data regulations in the US and China, for example, saying that *"You cannot use data collected in the US from a computer in China, and the other way also"*, and **I6** has underlined that *"We have for example, customers overseas and right now we know what we can share with them and so on, but it might change, you never know. So we take it as it comes and do the necessary actions when we need to do them; and it's also about what we need to share with the customers as well."* Another notable issue mentioned is the overall geopolitical state of the world, such as ongoing wars and conflicts affecting what, how and with whom their data can be shared.

4) Cost of Data: Data-related activities have direct implications for cost to the company. Data collection, analysis, storage and sharing – they all incur costs to the company. Hence, companies tend to assess business-related viability, such as generating Return on Investment (ROI) from any of their data-related activities, including data sharing. We received circumstantial evidence from our interviews regarding this. For example, **I4** has mentioned, *"I am a business guy, so my frustration was more around how do we use the data in order to do clever stuff to help the customers and for the company to make money on it?"*. This gives a clear indication of the expectations of the companies from data sharing activities. After all, data sharing is not considered a charity, similar to any other activity in a company.

I10 mentioned that the cost of data also includes the physical or virtual storage for this data, both of which are expensive. According to **I10**, *"In our case, only one hour of driving data equals one petabytes of data that includes the car's location, driving behavior, weather details, pictures and videos of the surroundings. It's too expensive to keep that. And sometimes the sensors might be out of sync or there could be dirt on it or something. So basically it's not useful information most of the time."* A remedy for this challenge can be training the AI and the ML systems to only collect insights regarding unique operational conditions and discard the rest.

In conclusion to the challenges section, the majority of the interviewees' concerns have been related to privacy of their data and the complexity of maintaining privacy through the different stages of data sharing. Different practices of overcoming the legal obstacles are needed to be investigated in order to make a comparison and more detailed conclusions.

V. Discussion

A. Contrasting the lines of a B2B data ecosystem

All the case organizations illustrate the importance of joint data management practices in terms of sharing and collaborating on data with their partners and customers, similar to other reports by, e.g., Holmström and Bosch [7]. The inter-organizational business relations underpinning the sharing create what we refer to as B2B data ecosystems, where data is collected, processed, and passed on through the focal organization. The organization of **I1**, e.g., collects diagnostics data from their video surveillance devices, which is used to identify potential bugs. In cases where this relates to a third-party application, such information is passed to enable a rapid solution.

Using the conceptual model by Runeson et al. [4], all of the seven case organizations may be characterized as *orchestrators*, or *platform providers*, of their respective data ecosystem – all with an organization-centric setup where the orchestrating organization decides what data is shared, and how. In contrast to an Open Data Ecosystem as those studied by Runeson et al., data is kept closed within the ecosystem and is regulated through agreements that include practices commonly used by similar organizations [19].

The type of data shared is driven and motivated by the business needs of the orchestrator and, by extension, its partners, customers, and other potential actors of the data ecosystem. Understanding this sharing and these relations in the context of business models was not explicitly investigated by this first study, but a fruitful topic for future research, along the lines of, e.g., Chakrabarti et al. [20].

Due to the sensitivity and business criticality of the data, many of the case organizations raised concerns about the security and integrity of the data collected and shared in their ecosystems. Encryption and sovereignty of data pipelines and platforms for sharing the data is therefore of critical concern for future infrastructures to support technological progress, and also an important area for future research, e.g., along the lines of Altendeitering et al. [21].

Regarding the amount of data shared, the interviewed organizations generally leaned towards collecting as much data as possible within the legal frameworks that apply, with a common rationale that AI-enabled systems require extensive

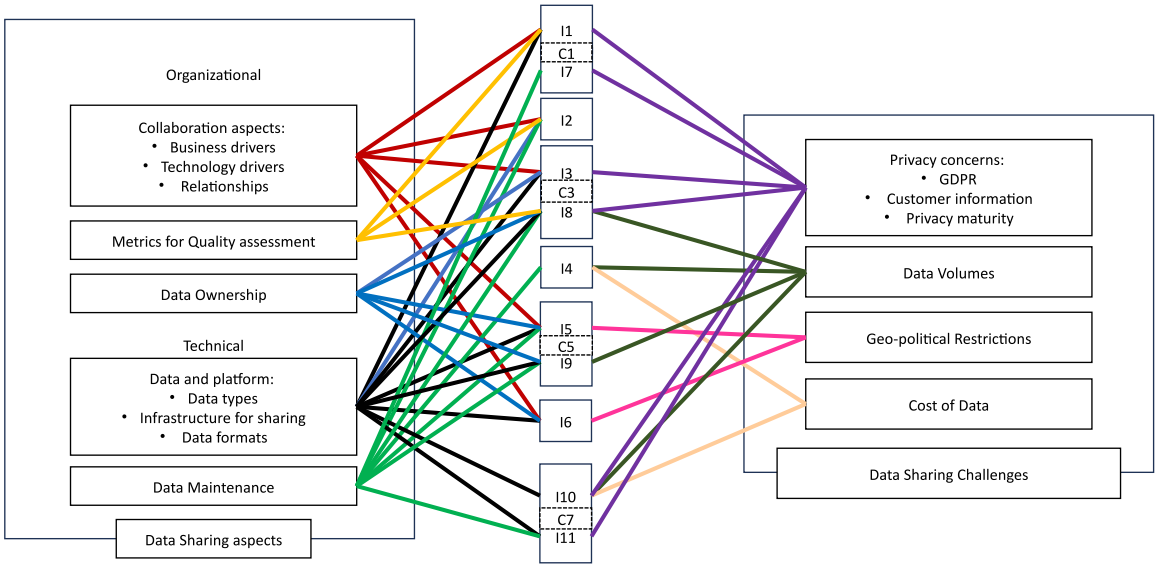

Fig. 3. Graphical representation of connections between the findings and interviews in this study.

training sets. This standing, however, comes with costs and challenges highlighted in the literature, e.g., relating to annotation, quality assurance, and storage [22].

B. Plausibility of generalization for a data sharing ecosystem

The interviewees have all presented evidence of importance of data sharing in their business. A uniform data sharing ecosystem solution that could allow to quickly add new members, perform various operations with a large variety of data types would potentially become a big enabler for the data-related developments. However, how plausible would such framework be, or is it just a utopia?

Legal challenges would be a large obstacle to consider when designing such a model. With respect to both internal and external regulations, the ability of adding such regulations into the system has to be taken into consideration, but how feasible is that? For that, data has to be intelligently split into subcategories, as regulations can differ from field to field, but are there cases that could cause legal issues just due to the complexity of the data nature from the legal standpoint? That is a question that would require a lot of research.

On the other hand, the technical requirements should be easier to handle. Ability to work with different data formats is already present in multiple solutions, and processing algorithms would potentially be added by the users, if not performed outside of the model. Volumes, on the other hand, could be an issue that should not be overlooked, as speed and storage capacity would define the efficiency of the data-related business processes.

However, at the current stage of the study it became apparent that the technology solutions that the companies are using are embedded into their workflow, especially when internal data sharing is being taken into consideration. For generalization, solutions need to be fully separable from the rest of the system in order to build them into the existing data sharing model, and our observation sets an important question of defining the requirements of the technologies in order to fulfil the potential research goal.

This being said, even these eleven interviews have presented a high number of different techniques and technologies for data sharing handling, both local or external (with, for example, Google and Amazon ecosystems being listed as such by **I5**), meaning that the uniform solution for all the data sharing needs will anyways be a tough task to tackle and should initially be tailored to the needs and technical possibilities of a company.

VI. Threats to validity

The study is an exploratory analysis of the phenomenon of intra-organizational data sharing. As the qualitative survey relies on the interviews with a small selection of participants, we make no strong claims about the findings. We have prioritized on *external validity* by selecting interviewees from seven organizations, rather than multiple representatives from one organization. Consequently, this reduces the *internal validity* as we heavily depend on single person's perspectives and opinions to represent one organization. A particular threat would be the lack of the company representatives holding strictly technical positions. *External validity* is threatened by the interviews being held only with the employees of companies based in the same country. On the other hand, all companies operate on an international market, and most of them with sites in multiple countries. *External validity* is also impacted by all the companies already operating in the area of data sharing, although we have purposefully selected such companies in order to investigate the existing practices and

challenges. The *construct validity*, i.e. to what degree there is a unified view on constructs and definitions of data sharing, is considered sufficient for this exploratory study, while for the continued study, a refined understanding of different types of data might be needed to explain the phenomena. The *reliability* of the data collection and analysis is considered good, as the researchers take turns in interviewing and analysis, as described in Section III-C.

This being said, as this is an exploratory analysis study, we do not consider these threats crucial, and will keep those in mind when conducting the future research.

VII. Conclusions and further work

We report the findings from a qualitative survey, comprising eleven interviews with representatives of data-driven companies in the context of data sharing. Based on our observations, data sharing is considered to be an important driver for businesses, with emerging technologies boosting capabilities of the companies.

With respect to the technical data, we observed a variety of technologies being used for the different methods of collecting, storing, processing and maintaining the data. While the topic of data sharing might still be fresh, the power of analysis tools such as machine learning (ML) leads companies towards developments in the area.

Throughout the coding synthesis, we found out the key challenges related to the data sharing processes. While being mostly related to the privacy matters, and while there are some internal practices established, usually in a way of contracts and user agreements, external regulations are an important area of interest and need to be investigated thoroughly in order not to fall in a legal trap.

As this is an exploratory study, its goal was to serve as a base for the future, larger study. The potential future research would be to conduct more interviews with a variety of different industry representatives in order to perform a full-scale multiple-case study [16], in order to come to a potentially more general view on how to design an easily extendable data sharing framework that could be used universally, which could serve as a theoretical base for a data sharing platform, solving the industrial data sharing needs.

Acknowledgement

Funded by the strategic research area ELLIIT (Excellence Center at Linköping–Lund in Information Technology) project C09, B2B Data Sharing for Industry 4.0 Machine Learning.

References

[1] B. Tronvoll, A. Sklyar, D. Sörhammar, and C. Kowalkowski, "Transformational shifts through digital servitization," *Industrial Marketing Management*, vol. 89, pp. 293–305, 2020.

[2] D. Coyle, S. Diepeveen, and J. Wdowin, "The value of data summary report," The Bennett Institute, Cambridge, Tech. Rep., 2020. [Online]. Available: https://www.bennettinstitute.cam.ac.uk/publications/value-data-summary-report/

[3] M. I. S. Oliveira, G. d. F. B. Lima, and B. F. Lóscio, "Investigations into data ecosystems: a systematic mapping study," *Knowledge and Information Systems*, pp. 1–42, 2019.

[4] P. Runeson, T. Olsson, and J. Linåker, "Open data ecosystems—an empirical investigation into an emerging industry collaboration concept," *Journal of Systems and Software*, vol. 182, p. 111088, 2021.

[5] J. Linåker and P. Runeson, "How to enable collaboration in open government data ecosystems: A public platform provider's perspective," *JeDEM - eJournal of eDemocracy and Open Government*, vol. 13, no. 1, pp. 1–30, 2021.

[6] W. Elgarah, N. Falaleeva, C. C. Saunders, V. Ilie, J. T. Shim, and J. F. Courtney, "Data exchange in interorganizational relationships: Review through multiple conceptual lenses," *SIGMIS Database*, vol. 36, no. 1, p. 8–29, feb 2005.

[7] H. Olsson and J. Bosch, "What got you here won't get you there. a multi-case study on the challenges in the transition from traditional towards continuous data practices in the embedded systems domain," in *1st International Conference on Software Product Management*. Gesellschaft für Informatik e.V., 2023.

[8] A. Aaltonen, C. Alaimo, E. Parmiggiani, M. Stelmaszak, S. Jarvenpaa, J. Kallinikos, and E. Monteiro, "What is missing from research on data in information systems? insights from the inaugural workshop on data research," *Communications of the Association for Information Systems*, vol. 53, pp. pp–pp, 08 2023.

[9] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "Data scientists in software teams: State of the art and challenges," *IEEE Transactions on Software Engineering*, vol. 44, no. 11, pp. 1024–1038, 2017.

[10] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, vol. 191, p. 111359, 2022.

[11] K. Hüner, A. Schierning, B. Otto, and H. Oesterle, "Product data quality in supply chains: the case of Beiersdorf," *Electronic Markets*, vol. 21, pp. 141–154, 2011.

[12] J. Gelhaar, P. Müller, N. Bergmann, and R. Dogan, "Motives and incentives for data sharing in industrial data ecosystems: An explorative single case study," in *56th Hawaii International Conference on System Sciences, HICSS*, T. X. Bui, Ed. ScholarSpace, 2023, pp. 3705–3714. [Online]. Available: https://hdl.handle.net/10125/103085

[13] M. Brechtel, D. Petrik, and K. Hölzle, "From challenges to solution pathways for industrial data ecosystems – A socio-technical perspective," in *Digital Responsibility: Social, Ethical, Ecological Implications of IS, 18. Internationale Tagung Wirtschaftsinformatik (WI)*. Paderborn, Germany: AISeL, 2023, p. 48.

[14] R. D'Hauwers and N. Walravens, "Do you trust me? value and governance in data sharing business models," in *Proceedings of Sixth International Congress on Information and Communication Technology*, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Singapore: Springer, 2022, pp. 217–225.

[15] P. Ralph, "ACM SIGSOFT empirical standards released," *SIGSOFT Softw. Eng. Notes*, vol. 46, no. 1, p. 19, 2021.

[16] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to advanced empirical software eng.* Springer, 2008, pp. 285–311.

[17] P. E. Strandberg, "Ethical interviews in software engineering," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM*. IEEE, 2019, pp. 1–11.

[18] J. Saldaña, *The Coding Manual for Qualitative Researchers. 3. edition.* Sage Publications, 2015.

[19] A. Dakkak, H. Zhang, D. I. Mattos, J. Bosch, and H. H. Olsson, "Towards continuous data collection from in-service products: Exploring the relation between data dimensions and collection challenges," in *28th Asia-Pacific Software Engineering Conf. (APSEC)*, 2021, pp. 243–252.

[20] A. Chakrabarti, C. Quix, S. Geisler, J. Pullmann, A. Khromov, and M. Jarke, "Goal-oriented modelling of relations and dependencies in data marketplaces." in *11th International i* Workshop co-located with the 30th CAiSE*, 2018. [Online]. Available: https://ceur-ws.org/Vol-2118/iStar2018_paper_4.pdf

[21] M. Altendeitering, J. Pampus, F. Larrinaga, J. Legaristi, and F. Howar, "Data sovereignty for AI pipelines: Lessons learned from an industrial project at mondragon corporation," in *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, ser. CAIN '22. New York, NY, USA: ACM, 2022, p. 193–204.

[22] D. Muiruri, L. E. Lwakatare, J. K. Nurminen, and T. Mikkonen, "Practices and infrastructures for machine learning systems: An interview study in Finnish organizations," *Computer*, vol. 55, no. 6, pp. 18–29, Jun. 2022.