Practices for Data Sharing: An Empirical Survey

Andrei-Raoul Morariu, Bogdan Iancu, Jerker Björkqvist
Åbo Akademi University
Faculty of Science and Engineering
Vesilinnantie 3, Turku, Finland
Email: {firstname.lastname@abo.fi}

***Abstract*—The data economy is changing the way companies operate. The largest companies in the world are strong actors in the data economy. Still, the share of data economy of GDP (Gross Domestic Product) is rather small. For industrial operators, data utilization is predicted to increase performance, predictability, and cost-effectiveness. However, to achieve the goals, data often need to be shared between operators, to produce system level gains. This paper analyses the possibilities and barriers for reaching effective data sharing through qualitative interviews with company representatives with technical insights into data sharing. The paper includes aspects such as value proposition, barriers, confidentiality, and technical aspects of data sharing.**

***Index Terms*—data economy, data sharing, industrial operators.**

I. Introduction

The volume of data/information created, captured, copied, and consumed worldwide was in the year 2010 approximately two zettabytes. This volume is forecasted to rise to 181 zettabytes by 2025 [1]. Most of the data is from people's activities where they are happy to share some of their data to improve their experiences. However, companies are more reluctant to share some of their data with other partners [2]. It is usually reduced towards data ownership using policies to prevent harming the participant entities [3].

This motivator made us proceed with a series of interviews with Finnish industry professionals to create an overview of the practicalities of data sharing. The data we refer to is not describing internet traffic data but measurement data from sensors. In most cases, such sensors are usually placed on moving mechanisms or adjacent to them collecting, e.g., vibration measurements.

Companies use data for various purposes, most prominent of which are: supplying it either as a standalone asset or a fundamental component of a product, improving their operational processes, or acquiring new technological or business insights. Data-driven innovation came to the forefront of modern industrial development in the past decade, since various technological advances stimulated data usage through significantly elevated storage capacity, computing power and data transmission speed [4] [5].

Data represents a crucial resource that is yet to be adequately exploited. Modern industrial development already employs prevalent data-intensive technologies such as machine learning, pervasive computing, edge computing, etc. [6]. One critical reason for the adoption of technologies that demand substantial data intake is the development of digital twins when simulating certain processes on physical models is of uttermost complexity. Data generation from diverse machines and devices contributes to datasets which can reach even petabytes per dataset. Companies, especially SMEs (Small and Medium Enterprises), that are reluctant to digitalization or even impede it, compromise their future progress relative to their competitors [4].

While many SMEs have the capacity to design highly innovative solutions to drive their business, in practice often they require data sources they do not have access to. Data sources are developed and analyzed by many companies primarily to advance their own business without an explicit intention to share it. Many data sources remain unexploited to the full of their potential even by the companies that produced them. Moreover, great reluctance towards sharing data with external partners seems to have permeated deep within business development layers as less than 50% of companies in a recent survey show that data sharing is a common practice in their company [7] [6].

In many companies, reluctance towards data sharing is rooted unsurprisingly in the economical aspect. A 2019 OECD (Organisation for Economic Co-operation and Development) report on data sharing identified the main impediments for accessing, sharing and exploiting data through continuous development as data privacy and ownership. Some states do not have currently any clear law that defines data ownership and the benefit of sharing [8]. The complexity is grounded in determining the requirements that need to be met prior to sharing data. Industrial players also find it challenging to estimate the value of data and to assess the risk of sharing it, with many of them being open to share data only if other players reciprocate [6]. Another critical aspect that hinders data sharing among companies is inadequate transparency and a stark imbalance in power between different players in a market sector [5]. One of the most prevalent challenges for data sharing is privacy, which affects several aspects of development within an ecosystem: maintenance of data through its entire life-cycle [9], safeguarding it from corruption, and sustainable development to retain usability [4].

B2B (Busines-to-Business) data sharing is especially scarce, impeding data utilization to its full potential at industrial level [5]. There is a great discrepancy between data sharing and the extensive efforts that are put into collecting it [10].

 ISBN: 978-1-61208-994-2

Notable benefits were attained in some industrial sectors through data sharing. One example is the automotive industry and transport sector, where data sharing promoted smart mapping solutions. For instance, the mapping service HERE [11] provides enhanced spatial data, including geocoding, positioning, map rendering etc., and is endorsed by some of the most prominent automobile manufacturers. While increasing their revenue, HERE also is committed to contribute to safer and more sustainable transportation. Furthermore, automobile manufacturers can capitalize on their data through the HERE marketplace [10]. Another sector where data sharing has brought about tremendous benefits is the healthcare sector, where data sharing is used to promote more rapid diagnostics and more effective treatments by employing machine learning algorithms [12] [13]. Not only does medical data sharing benefit society at large by providing enhanced medical services, but medical providers also capitalize on these solutions, offering more efficient and precise diagnostics [14] [15].

There are two distinct data sharing strategies when it comes to private data, *vertical*, which develops across the supply chain, and *horizontal*, which transpires between competing companies [10] [16]. Vertical data sharing is characterized predominantly by trust between companies along the supply chain and the degree of certainty attributed to certain business needs. Horizontal data sharing, however, is employed to a lower degree due to its sensitivity. Personal data, which concerns largely horizontal data sharing is highly regulated in the European Union by GDPR (General Data Protection Regulation), which requires extensive considerations on various aspects of its content. This, in turn, makes personal data sharing a very intricate matter [10].

To promote data sharing by various companies, it becomes imperious to incentivise the process since it involves an extremely valuable asset. One way to do so could be to purchase it. However, data acquisition raises yet new challenges since selling it at an equitable price can be a difficult endeavour [17] [18]. To address such demands, data marketplaces emerged, such as: Azure Marketplace [19], Japan Data Exchange Inc. [20], Qlik DataMarket [21], etc. However, a considerable variety of data precludes trading it in a fairly regulated manner [22]. To address these challenges, different pricing and trading models showcasing auspicious prospects were introduced in [23] [24].

The paper is organized as follows. Section II depicts the study design and interview questions. Section III illustrates the results of the qualitative analysis of interviews with companies. In Section IV, a few solutions for data exchange are discussed. Conclusions of the study are presented in Section V.

II. Research Questions and Study Design

Sharing data is an integral part of the scientific community as it allows for verification of results and enables researchers to build upon previously discovered information [25]. Interest in this subject came from multiple face-to-face conversations with industry experts. They assert that cooperation between companies on similar topics can lead to further developments in different areas connected to the industries. Cooperation under non-disclosure agreements would secure the data transactions and enforce the credibility between partners.

The study addresses the following research questions:

RQ1: Is your organisation exercising any data sharing with other company?
  a) If yes, to what extent do you do this and how dependent is your business on data sharing? How long do you store the data you share with others? Do you use a third part service for sharing your data, such as Amazon/cloud services?
  b) If no, what are the reasons for not doing this?

RQ2: What value do you see between sharing your data with other organisations?

RQ3: What are the barriers for performing data sharing? Were there some previous episodes of data sharing with others?

RQ4: Is confidentiality of data an issue?

RQ5: Do you have a data management strategy/policy?

Research questions were used to extract data from a series of interviews with experienced employees from different business areas. We organized Semi-Structured Interviews (SSI) for our research. A semi-structured interview is a meeting where the interviewer asks questions that are meant to evoke open conversations offering participants the chance of bringing up important matters [26]. As part of their experience, SSIs are intended to find out what participants think about their experiences related to the given topic [27]. It provides the advantage of collecting the data needed from topics that can be further developed towards new insights.

A. Threats to validity

The most significant threat to the validity of RQ1 is that data sharing between companies is still limited. Furthermore, companies' guidelines and restrictions drive many respondents to give short and sparse answers to questions. The purpose of RQ1 is to develop a more in-depth understanding of data sharing between companies and other partners. Using this question, we learn how vital data provided by other partners is to a business's success, such as data availability and the platform used for data sharing.

Data value is a subjective opinion that allows for RQ2 to bear a respective degree of uncertainty upon receiving an answer from participants. Although, considering the expertise of the interviewees, the article may provide inspiration for other businesses.

The purpose of the RQ3 is to challenge the specialist to observe the main topic from a different point of view. This question may also bring challenges allowing the respondent to claim that the company is conducting safe data sharing procedures with other external partners. Moreover, RQ3 further investigates whether there have been previous agreements on sharing data with other partners and the resulted outcome. In this case, common factors such as lack of time or limited

 ISBN: 978-1-61208-994-2

funding prevent companies from sharing their data with other external partners [25].

RQ4 is counted as a controversial question where many respondents had to take some extra time to respond. Many thought that confidentiality could lead to security concerns since raw data does not display any signs of being private. On the other hand, even number sequences can include, for example, personal identification codes from a specific country.

The validity of our study of RQ5 is compromised by companies keeping the response to the question private. Therefore, companies could decide on some data management strategies internally.

B. Conducting the interview

The selection of respondents was decided internally within the author group on companies that were previous project members in consortiums with the university. Considering their experience and previous connection to us, we interviewed the selected specialists.

When we conducted the interview, the concept of sharing data with other partners was still not a regular practice. Many specialists responded negatively to whether they share company data with external parties. A quarter of the total number of specialists invited for interviews chose not to attend online interviews due to their personal disinterest in discussing any information about company practicalities.

A total of 12 specialists were interviewed during three weeks between 2022-04-02 to 2022-04-28. In selecting the number, we focused on gathering experts from diverse organizations and fields. Interview meetings were agreed upon according to the availability of the specialists. In an email before the interviews, participants were informed that the study would lead to a publication that would enhance university research and provide new business opportunities for companies. Each interview was scheduled for 30 minutes, with an average of 24 minutes per participant. There was an interview guide containing the research questions used listed in Section II. Many questions were meant for an open discussion where the specialists discussed the practicalities of data sharing within the companies where they work. The interviews were transcribed during the conversation in real-time.

III. Results

In this section, we present the main findings of the research. The interview results where categorised in a systematic way based on each question and follow-up discussions.

We aim to create a comprehensive guideline for data sharing that would encourage companies to increase their expertise and revenue from collaborations.

The research starts with domain identification where the respondents were working at the interview date. Figure 1 illustrates these domains. We chose the working domains of the companies with generic names to keep them transparent. We focused on having different areas of expertise of the respondents, but only with applicability to the industry. We chose an even distribution of chart slices for company domains to increase the privacy of the respondents. Several employees working at the same place were interviewed for some of the companies. We decided this way because the experience of the interviewed persons was valuable to our research. Furthermore, to minimize the repeatability of answers, we only chose respondents from different departments of the same company.

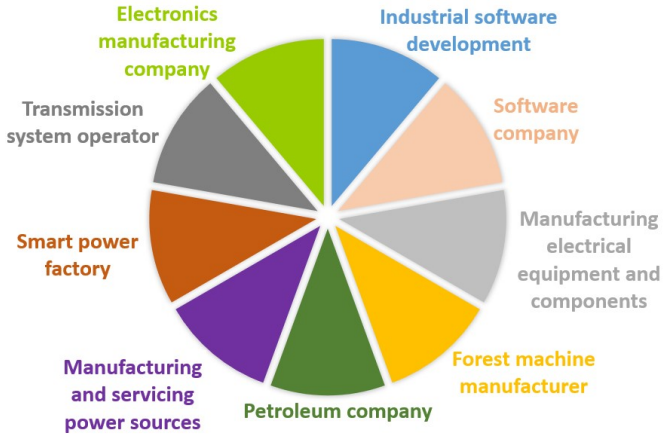

Fig. 1. Application domains of the company representatives who participated in the qualitative interviews.

A. Data sharing between companies (RQ1)

Most of the time, sharing data is hindered by company culture and practices. In addition, individual judgment plays a significant role in influencing later decisions. Today, new mandates relate to data management strategies, which means that new partnerships will be more open.

Considering the first research question, the respondent should share with the researchers whether the company where they work shares any data with outside partners. There are some follow-up questions on the given answers. The purpose of this question is to collect information on reasons for not sharing data or on how dependent the business is on distributing data.

Starting with this, some of the experts responded that sharing data may lead to: further improvements in manufacturing and execution, improve analytics or maintain safe operations. Here, we also asked on how was this performed. The most common answer noted was via project partnership, while others mentioned about verbal agreement, or one time transactions.

It was discussed that confidentiality, general data protection regulation, data ownership, and low business opportunities were some of the risks preventing data sharing. As one individual indicated, their business was not dependent on data sharing. Therefore, it was not taking place. Yet another claimed that data sharing had been attempted with another partner. This practice did not result in a long-term partnership due to the differences between the two partners' technologies.

In addition to the opening question, we asked how long data shared with others is stored. The most common answer

 ISBN: 978-1-61208-994-2

is that the data is erased according to customer agreements and when the project ends. A few respondents noted that they typically archive the data for up to 10 years. They added that some reasons involve legal requirements and testing information (e.g., electrical components). Furthermore, some mentioned that they prefer storing data for a longer duration for availability on further improvements to their products and services.

Many respondents noted that they use an in-house private cloud for data sharing. According to Figure 2, Google, Amazon, Microsoft Azure, and OneDrive are services used for data sharing. Others mentioned file transfer protocol (TCP/IP) and email attachments as means for data sharing. This question also brought up an emerging topic from one of the respondents. A respondent stated that they are involved with creating a data lake where companies would be able to access data according to agreements. This illustrates the need for industry companies to develop data sharing frameworks.

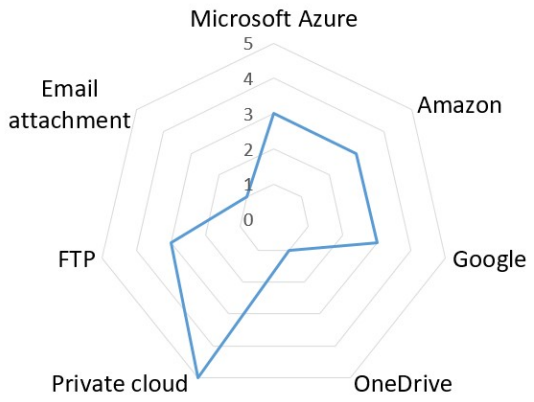

Fig. 2. Means of performing data sharing with external partners.

B. Value of shared data (RQ2)

As mentioned in Section II-A, the answer to this question differs even for people from the same company. It aims at experts' opinions on data visualization, which sometimes may only mean a series of numbers aligned in a specific order.

Experts' viewpoint is that the value of shared data comes from reasons such as:

- Money saving from better understanding of the processes
- Forecasting the need for components
- Quick troubleshooting
- Enhancing services to customers
- Development of products
- Increasing value of use cases
- Ensuring customers that sold products will not break

A few interviewees noted that perhaps the value in sharing data with other partners is yet to be discovered or even speculated that there might not be value in doing it at all. They added that the data recipients could have additional benefits from analyzing it and that pre-processing it for sharing purposes typically requires additional resources. Furthermore, sometimes the data the company is working with may belong to clients and is, therefore, more complicated to share externally. In some cases, data refers to the information given via a telephone call from one person to another. Such a situation leads to miscommunication with the other groups interested in the details.

C. Barriers from performing data sharing (RQ3)

Obstacles preventing data sharing can sometimes be tied to cultural reasons preventing people from making such decisions. Data sharing is hampered by intellectual property rights due to ownership of data. Some companies are service providers using data from partners according to contracts. In this case, sharing customer data would mean creating new agreements, contacting the customer, and other connected processes.

In Figure 3, we extracted some of the reasons preventing companies from exercising data sharing. In many cases, experts contend that the heavy process, confidentiality, and Non-Disclosure Agreement (NDA) are some of the main reasons for not executing data sharing. Usually, it takes a long time to agree on the revised terms of the contract with customers, and the process becomes extremely tedious. It is crucial to consider privacy and access when it comes to data sharing to prevent breaches, identity theft, or other security threats.

Therefore, some respondents mentioned that before sharing the data with others, they perform data masking in various ways such as cleaning, averaging, anonymizing, and removing the means. There is no harm intended to occur to the company sharing the data through those filtering methods.

Fig. 3. Barriers preventing data sharing between companies.

D. Is data confidentiality an issue? (RQ4)

Having information about a company's process work may be of significant importance to the company. However, that same piece of information may not be of any importance to another company. This leads to a comprehensive definition of confidentiality. This interview question raised many affirmative answers on validating that data confidentiality is an issue. Experts mentioned that data sharing is always executed under NDA contracts and trust. The NDA contracts usually contain agreements that include:

- Common collaboration on improvements

 ISBN: 978-1-61208-994-2

- Complete list of information about shared data
- Data leaking prevention measures
- Access limitations

E. Data management strategies/policies (RQ5)

Three-quarters of the respondents mentioned that the company where they work has data management strategies. Distributed control systems and in-house data management are the most common strategies related to cyber security. Data systems in many companies are only used within a closed network without an external data sharing interface.

Equipment manufacturers typically create an Application Programming Interface (API) to facilitate customers' installation of their protocols. One respondent mentioned that the data management policy implies retrieving data up to ten years old.

IV. Strategies for performing data exchange

Machine learning models that use the output to change the operation of a real-world device may be patentable since they are integrated into changing the actual state of the device. Artificial intelligence systems are now able to solve problems more effectively due to new developments. But a novel solution may still be patentable if it improves upon a conventional manual process as well. Last but not least, it is crucial for companies to prepare detailed disclosures describing the low-level details of the system [28].

Two core challenges affect the deployment of AI solutions in companies. Firstly, industrial partners store data in silos and do not implement regulated exchange frameworks, and secondly they prefer a more traditional approach to data privacy and security, which limits considerably data sharing or exchange, sometimes to an astounding degree. Given the highly competitive nature of various industrial sectors, enhanced privacy requirements and demanding data management specifications, companies find it often challenging to implement data integration even within their own organisation. One approach to address the aforementioned hindrances is *federated learning*, which promotes the idea of developing Machine Learning (ML) models by exploiting data originating from multiple sources, while obstructing data breaches [29].

An important challenge that federated learning faces is the *traceability* of ML models throughout their life-cycle. Given a prediction value from an ML process, if one cannot trace which input values (originating in which datasets) determined it, we encounter a situation when the ML model is essentially a black-box, hence its traceability cannot be ensured [30]. To address this topic, several frameworks employing blockchain technology emerged, which tackle the development of more transparent deep learning models with enhanced traceability [31] [32].

Gaia-X [33] is a standard for data exchange across companies. Its role is to be a mediator in agreements between companies and to ensure cooperation. One of these data ecosystems is Catena-X [34], which is responsible for creating a standalone data exchange standard across the entire automotive supply chain. The core values of their standards are to ensure data protection, security, and fairness for participating companies.

Toolchains have been standardized by numerous non-profit organisations, such as ASAM (Association for Standardization of Automation and Measuring Systems), to ensure better quality for their underlying processes, testing and development in the automotive sector [35]. The members of ASAM standards are companies involved in the car manufacturing process as manufacturers, suppliers, etc. ASAM is the owner of the standards that enable data exchange or the necessary tools required. In order to share data within the ASAM group, partners need to comply with the definition of the test data provided by the application model and to have the data in XML (Extensible Markup Language) file format. Since every company has its application model, the ASAM standard extends toward company-specific metadata [36].

V. Conclusion

Data economy is an emerging topic where many companies are currently working to improve their operability and increase revenue through the development of various systems.

Many companies and allied businesses begin to exchange data in order to increase the number of services they offer and solve unknown customer problems.

This article aims to provide an overview of various companies' capabilities to collaborate with external partners across multiple sectors. We interviewed 12 employees of various companies in industrial sectors that gave us insights on practices they employ regarding data sharing with external parties.

It is common nowadays to see an increase in external collaboration, but unfortunately, companies are backed-up on collaborating only under projects. Collaboration between two companies is usually time-consuming when NDA contracts are involved. For the purpose of avoiding damaging company information, companies prefer to send single batches of data that are averaged and partially removed.

In order to increase the number of services offered, young emerging businesses must make their customers aware of the potential data sharing.

Acknowledgments

This work was partially supported by projects funded by Business Finland.

References

[1] Statista. https://www.statista.com/, Accessed on June 2022.

[2] N. Pearce and A. H. Smith. Data sharing: not as simple as it seems. *Environmental Health*, pp. 1–7, vol. 10(1), 2011.

[3] F. Aggestam. Setting the stage for a shared environmental information system. *Environmental Science & Policy*, pp. 124–132, vol. 92, 2019.

[4] S. Yin et al. Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics*, pp. 657–667, vol. 62(1), 2014.

[5] H. Richter and P. R. Slowinski. The data sharing economy: on the emergence of new intermediaries. *IIC-International Review of Intellectual Property and Competition Law*, pp. 4–29, vol. 50(1), 2019.

[6] A. Krotova, A. Mertens, and M. Scheufen. Open data and data sharing: An economic analysis, IW-policy paper. Technical Report 21, Institut der deutschen Wirtschaft (IW) Köln, 2020.

[7] H. Huttunen et al. What are the benefits of data sharing? uniting supply chain and platform economy perspectives. *Uniting Supply Chain and Platform Economy Perspectives (September 19, 2019)*, 2019.
[8] X. Li and Y. Cong. A systematic literature review of ethical challenges related to medical and public health data sharing in china. *Journal of Empirical Research on Human Research Ethics*, pp. 537–554, vol. 16(5), 2021.
[9] H. F. Atlam and G. B. Wills. Iot security, privacy, safety and ethics. In *Digital twin technologies and smart cities*, pp. 123–149. Springer, 2020.
[10] N. Stüdlein. Developing a framework for strategic data sharing barriers among competitors. *Data as a Common Good*, pp. 16, 2022.
[11] HERE. https://www.here.com, Accessed on August 2022.
[12] G. Cammarota et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews gastroenterology & hepatology*, pp. 635–648, vol. 17(10), 2020.
[13] K. Y. Ngiam and W. Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, pp. 262–273, vol. 20(5), 2019.
[14] S. Rutella et al. Society for immunotherapy of cancer clinical and biomarkers data sharing resource document: volume i—conceptual challenges. *Journal for immunotherapy of cancer*, vol. 8(2), 2020.
[15] D. Kerr et al. The oncology data network (odn): A collaborative european data-sharing platform to inform cancer care. *The Oncologist*, pp. 1–4, vol. 25(1), 2020.
[16] F. Cruijssen, W. Dullaert, and H. Fleuren. Horizontal cooperation in transport and logistics: a literature review. *Transportation journal*, pp. 22–39, vol. 46(3), 2007.
[17] D. Iwasa, T. Hayashi, and Y. Ohsawa. Development and evaluation of a new platform for accelerating cross-domain data exchange and cooperation. *New Generation Computing*, pp. 65–96, vol. 38(1), 2020.
[18] M. Zhang et al. Pricing fresh data. *IEEE Journal on Selected Areas in Communications*, pp. 1211–1225, vol. 39(5), 2021.
[19] Azure Marketplace. https://azuremarketplace.microsoft.com/en-u, Accessed on July 2022.
[20] J-DEX. https://j-dex.co.jp/en/index.html, Accessed on July 2022.
[21] Qlik. https://www.qlik.com/us/, Accessed on July 2022.
[22] F. Liang et al. A survey on big data market: Pricing, trading and protection. *IEEE Access*, pp. 15132–15154, vol. 6, 2018.
[23] J. Yang, C. Zhao, and C. Xing. Big data market optimization pricing model based on data quality. *Complexity*, 2019.
[24] Z. Zheng et al. Arete: On designing joint online pricing and reward sharing mechanisms for mobile data markets. *IEEE Transactions on Mobile Computing*, pp. 769–787, vol. 19(4), 2019.
[25] C. Tenopir et al. Data sharing by scientists: practices and perceptions. *PloS one*, pp. e21101, vol. 6(6), 2011.
[26] C. Schmidt. The analysis of semi-structured interviews. *A companion to qualitative research*, pp. 7619–7374, vol. 253(258), 2004.
[27] M. J. McIntosh and J. M. Morse. Situating and constructing diversity in semi-structured interviews. *Global qualitative nursing research*, pp. 2333393615597674, vol. 2, 2015.
[28] E. L. Sophir and R. E Glass. Key strategies for patenting big data solutions. https://www.foley.com/en/insights/publications/2022/key-strategies-for-patenting-big-data-solutions, Accessed on July 7 2022.
[29] Q. Yang et al. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, pp. 1–19, vol. 10(2), Jan 2019.
[30] V. Mothukuri et al. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, pp. 619–640, vol. 115, 2021.
[31] H. Kim et al. Blockchained on-device federated learning. *IEEE Communications Letters*, pp. 1279–1283, vol. 24(6), 2019.
[32] K. Salah et al. Blockchain for ai: Review and open research challenges. *IEEE Access*, pp. 10127–10149, vol. 7, 2019.
[33] Gaia-X. https://gaia-x.eu/, Accessed on July 2022.
[34] Catena-X. https://catena-x.net/en/, Accessed on July 2022.
[35] ASAM. https://www.asam.net/, Accessed on July 2022.
[36] ASAM. Asam solutions guide. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.738.2707rep=rep1type=pdf, Accessed on July 10 2022.

 ISBN: 978-1-61208-994-2