Modeling and Operationalizing Data Ecosystems

Soo-Yon Kim
Supervised by: Sandra Geisler
RWTH Aachen University
Aachen, Germany
kim@dbis.rwth-aachen.de

ABSTRACT

Data sharing and reuse are increasingly recognized as key drivers of innovation, efficiency, and collaboration across domains. Data ecosystems (DEs) provide a framework for enabling such value creation, but their inherent complexity, spanning technical, organizational, and social dimensions, makes them difficult to understand and implement effectively.

This research addresses these challenges by developing a model that draws from a life cycle modeling approach and is informed by domain-specific experiences from supply chain and scientific research use cases. The model captures the key stages of data sharing in data ecosystems, such as data selection, preparation, and indexing. For each use case, the steps are described in terms of the involved tasks, roles, actors, requirements, and needs, enabling a structured view of how data sharing unfolds across technical and organizational boundaries in practice. This perspective aims to support both the understanding and purposeful design of data ecosystems with a specific focus on the processes they are intended to facilitate.

The application of the life cycle modeling approach will be evaluated based on its usefulness in analyzing and supporting real-world data sharing processes. Prototype tools will be developed to address specific, recurring challenges observed in selected stages of the data sharing life cycle, such as data preparation and analysis, and will be evaluated based on their relevance to practitioner needs and their contribution to the success of the respective life cycle stage.

Through this work, the project aims to respectively contribute to, and bridge the gap between, conceptual understanding and practical implementation, supporting the development of value-creating, scalable data ecosystems.

VLDB Workshop Reference Format:
Soo-Yon Kim and Supervised by: Sandra Geisler. Modeling and Operationalizing Data Ecosystems. VLDB 2025 Workshop: PhD Workshop.

1 INTRODUCTION

As the amount of data in organizations continues to grow, so does the volume of data that remains unused or siloed. There is an increasing interest in unsiloing, sharing, and (re-)using this data. Combining data from different sources, especially those not previously considered together, can lead to new insights, to improved products and processes, and to deeper and more nuanced knowledge.

The research of the Cluster of Excellence project "Internet of Production"[1] highlights two domains where relevant use cases of data sharing can be found.

The first is the domain of industrial production, which is the main focus of the "Internet of Production" project. We take supply chains as an example. Supply chains are susceptible to disruptions, as shown during the Covid-19 pandemic. The ability to respond more effectively to such events could be improved if companies shared real-time information about the status of deliveries and goods in transit [17].

The second domain is the academic context itself. Widely supported by research funders and institutions, there is a growing interest, and mandates, in making science and research data more open and reusable. When insights are shared, they can complement, validate, or add to each other, and thus deepen existing, or generate new, knowledge.

While the potential benefits and use cases of data sharing are rather apparent, establishing data sharing is not trivial in practice. Many questions arise: What are the incentives for sharing data for data providers? How can potential data users find out what kind of data is available for reuse? How can privacy and data ownership be preserved, especially for sensitvie data? How can companies share certain information without exposing business-critical details? What kind of infrastructure is required to handle the sharing of very large amounts of data? Is the data that is shared comprehensible for both humans and machines?

Such environments centered around data sharing have been coined in the literature with the term *data ecosystem (DE)* [14]. DEs are highly complex, entailing both social and technical components, which interact with and influence each other. To make DEs work effectively, we need a better understanding of DEs in general, and how they can be purposefully designed, adapted, and operated.

Current research on DEs follows, among others, the subsequent three directions [19]:

The first direction revolves around creating a better understanding of DEs. The works focus on developing theories and models. The second direction studies use cases, including value creation in DEs and exploring domain-specific characteristics. The third direction revolves around DE management and engineering, addressing topics such as security and data integration.

While valuable work exists in all three areas, there are still gaps. Some models are outdated or focus too narrowly on specific aspects. Generic approaches are often too abstract for real-world use,

Proceedings of the VLDB Endowment. ISSN 2150-8097.

[1] https://www.iop.rwth-aachen.de/cms/~gpfz/produktionstechnik/?lidx=1

while practical solutions sometimes lack generalizability or a solid conceptual foundation.

This PhD project addresses these gaps by applying a structured life cycle modeling approach to data sharing in data ecosystems, grounded in both modeling literature and real-world use cases from supply chains and scientific research. It focuses on concrete steps such as data selection, preparation, indexing, and exchange, along with the associated actors, their requirements, and the organizational and technical contexts. Based on this foundation, targeted tools will be developed to address challenges in these stages. The following section positions this approach within the existing body of research.

2 RELATED WORK

DE models. DEs are highly complex, socio-technical systems involving a wide range of actors, technologies, and organizational processes. Several modeling approaches have been proposed to better understand and structure this complexity. For example, [15] have introduced a meta-model that describes the core elements of DEs, including actors, roles, resources, and relationships. Gelhaar et al. have contributed to the conceptual understanding of DEs through multiple works: analyzing how DEs emerge [10], developing a taxonomy of DEs [9], and examining relationships in DEs [8]. The International Data Spaces Associaton[2] has drafted a reference architecture describing sovereign data exchange between industrial partners [16]. Game-theoretic models have also been explored to describe data pricing in data markets, modeling data as economic goods [3]. A recent work with a life cycle modeling approach describes socio-technical processes in terms of sequential steps, roles, and requirements [6]. Although originally developed for knowledge graph ecosystems, it offers a transferable perspective that may also be applied to processes in data ecosystems.

Applied perspectives on DEs. Various application areas reflect distinct requirements and dynamics. [1] and [4] have studied data sharing practices between businesses, highlighting both enablers and barriers to collaboration. In the context of public administration, [21] and [13] have examined how DEs can support data-driven governance and public administration. [5] have investigated the development of a mobility data space, focusing on the challenges of cross-provider data integration in transportation systems. [20] have contributed insights into the structure and governance of DEs in industrial settings.

DE management and engineering. Another important area of research addresses the enabling of DEs through appropriate management and the engineering of DE solutions. Several works research semantic interoperability [2, 22]. Other works focus on data integration [11] and analysis across distributed systems [22]. Further works address the need for cross-organizational transparency and data quality [7, 12]. Finally, [18] have examined DE governance from an ethical standpoint, considering issues such as fairness and accountability.

Despite the breadth of existing models and overviews of DEs, the models are too abstract to be applied to real-world scenarios, are not up-to-date, or focus only on certain aspects. Conversely, domain-specific studies sometimes lack broader applicability or theoretical grounding. Practical tools and methodologies that support the operationalization of DE concepts in concrete settings are still scarce. There is a need for research that provides actionable process models, grounded in representative use cases, capable of guiding the analysis, design, and implementation of key operational processes in data ecosystems.

[2]https://internationaldataspaces.org/

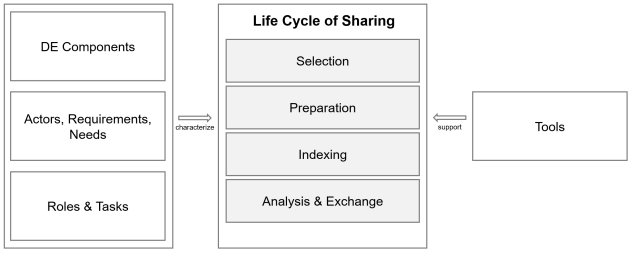

Figure 1: Data sharing life cycle with associated elements and tool support.

3 RESEARCH OBJECTIVES AND QUESTIONS

The objectives of the PhD project are structured around the following research questions.

RQ1: How can the steps and conditions of data sharing processes in data ecosystems be modeled to support their implementation?

This question focuses on adapting and applying an existing life cycle meta-model to describe the structure of data sharing processes. The aim is to identify typical steps (e.g., data selection, preparation, exchange), define their dependencies, and capture relevant actors, roles, tasks, and conditions under which these steps take place.

RQ2: What challenges arise in the data preparation stage, and how can they be addressed through targeted support?

Data preparation is a recurring bottleneck in data sharing. This question aims to analyze technical and organizational challenges observed in real-world settings and to develop tools that are customized for meeting the requirements of the involved actors and for effectively contributing to the success of this step.

RQ3: What challenges arise in the joint analysis and exchange stage, and how can they be addressed through targeted supported?

This question focuses on identifying the requirements, constraints, and coordination needs involved specifically in the analysis and exchange of data across organizations. Based on this understanding, tools or methods will be identified and developed to support the overall execution of this life cycle step.

Figure 1 summarizes the life cycle model, the associated elements, and the targeted tool support. The following section outlines the methodology used to address these research questions.

4 METHODOLOGY

The methodological approach is divided into two main parts: first, the application of a life cycle modeling framework to describe and analyze data sharing processes in selected use cases; second, using these models to guide the development of tools that address specific challenges within individual life cycle stages.

In the first part of the methodology, we apply an existing life cycle modeling framework [6]. While originally developed for knowledge

graph ecosystems, we adapt this framework to the context of DEs, focusing specifically on the life cycle of data sharing.

Modeling is carried out directly on two representative use cases: one from the domain of industrial supply chains, focusing on joint KPI calculation, and one from scientific research, focusing on data sharing for collaborative research. For each case, we identify the relevant steps in the data sharing process such as data selection, preparation, indexing, analysis, and exchange, and describe them using the elements provided by the life cycle framework. This structured representation supports the analysis of dependencies, bottlenecks, and coordination needs across actors and organizations.

For modeling the elements characterizing each step, we consider different modeling languages, such as UML or i* to model actors and their goals and needs, or Business Process Model and Notation for modeling roles and tasks.

In the second part of the methodology, the life cycle models developed for the two use cases are used to identify specific technical and organizational challenges in the data sharing process.

To prioritize areas for tool support, tasks within each step are examined with respect to their complexity, frequency, and relevance across both use cases. Particular attention is given to activities that require substantial manual effort, involve cross-organizational infrastructure, or depend on specialized knowledge. Based on this analysis, and following a design science research methodology, prototype tools are developed for selected steps in the data sharing life cycle. These include, for example, support for standardized data annotation, development of cross-organizational data catalogs, and distributed analysis.

The development of the tools follows an iterative process aimed at aligning functionality with the specific needs and constraints identified during the first part of the research.

The following section outlines the evaluation strategy used to assess the developed models and tools.

5 EVALUATION

The evaluation of this work is structured along three dimensions: applicability, actionability, and usefulness.

Applicability examines whether the life cycle model can be effectively applied to diverse use cases. The evaluation will assess whether the model captures relevant steps, roles, and dependencies in actual use cases. This will be conducted by applying the framework to the two use cases from the domains of supply chains and scientific research. It will be further explored whether the model can be adapted to other domains or life cycles such as data reuse.

Actionability assesses whether the modeling outputs provide practical guidance for system developers and decision-makers. Evaluation will focus on whether the models offer clear guidance for designing and organizing data sharing processes. This will be done through structured feedback sessions with domain experts and practitioners who will assess the clarity and usefulness of the models for guiding real decisions.

Usefulness evaluates the effectiveness of the developed tools in supporting selected life cycle steps. This includes their relevance to practitioner needs, their potential to reduce manual effort, and their support for coordination tasks. Tools will be tested through task-based evaluations and interviews with practitioners involved in the use cases.

Together, these evaluation activities aim to assess whether the modeling approach and tools are transferable, practically informative, and effective in supporting data sharing processes in real-world settings.

6 CONCLUSION

The research of this PhD project aims to advance the understanding and practical implementation of data ecosystems by applying a life cycle modeling approach to real-world data sharing use cases and developing tools to support selected process steps. By combining a use case–driven modeling approach with design science research, the project aims to contribute to advancement in both the conceptual understanding and practical operationalization of DEs. The planned evaluation ensures that the outcomes are applicable and transferable to different domains, provide a practical guide for DE implementation, and effectively support data sharing. Ultimately, the project seeks to contribute to more efficient, value-creating, and scalable data sharing practices across organizations.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

REFERENCES

[1] Antragama Ewa Abbas, Wirawan Agahari, Montijn van de Ven, Anneke Zuiderwijk, and Mark de Reuver. 2021. Business Data Sharing through Data Marketplaces: A Systematic Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research* 16, 7 (2021), 3321–3339. https://doi.org/10.3390/jtaer16070180

[2] Sebastian Bader, Jaroslav Pullmann, Christian Mader, Sebastian Tramp, Christoph Quix, Andreas W. Müller, Haydar Akyürek, Matthias Böckmann, Benedikt T. Imbusch, Johannes Lipp, Sandra Geisler, and Christoph Lange. 2020. The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. In *The Semantic Web – ISWC 2020*, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Lecture Notes in Computer Science, Vol. 12507. Springer International Publishing, Cham, 176–192. https://doi.org/10.1007/978-3-030-62466-8{_}12

[3] Yuran Bi, Yihang Wu, Jinfei Liu, Kui Ren, and Li Xiong. 2024. When Data Pricing Meets Non-Cooperative Game Theory. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5548–5559. https://doi.org/10.1109/ICDE60146.2024.00443

[4] Ruben D'Hauwers, Nils Walravens, and Pieter Ballon. 2022. Data Ecosystem Business Models. *Journal of Business Models* 10, 2 (2022), 1–30. https://doi.org/10.54337/jbm.v10i2.6946

[5] Holger Drees, Johannes Lipp, Sebastian Pretzsch, and Christoph Schlueter Langdon. 2021. Mobility data space–first implementation and business opportunities. In *ITS World Congress*.

[6] S. Geisler, C. Cappiello, I. Celino, D. Chaves-Fraga, Anastasia Dimou, A. Iglesias-Molina, M. Lenzerini, A. Rula, Dylan Van Assche, S. Welten, and M.-E. Vidal. 2025. From genesis to maturity : managing knowledge graph ecosystems through life cycles. *PROCEEDINGS OF THE VLDB ENDOWMENT* 18, 5 (2025), 1390–1397. https://doi.org/10.14778/3718057.3718067

[7] Sandra Geisler, Maria-Esther Vidal, Cinzia Cappiello, Bernadette Farias Lóscio, Avigdor Gal, Matthias Jarke, Maurizio Lenzerini, Paolo Missier, Boris Otto, Elda Paja, Barbara Pernici, and Jakob Rehof. 2022. Knowledge-Driven Data Ecosystems Toward Data Transparency. *Journal of Data and Information Quality* 14, 1 (2022), 1–12. https://doi.org/10.1145/3467022

[8] Joshua Gelhaar, Felix Becker, and Tobias Groß. 2022. Characterization of Relationships in Data Ecosystems. (2022). https://doi.org/10.15488/12159

[9] Joshua Gelhaar, Tobias Groß, and Boris Otto. 2021. A Taxonomy for Data Ecosystems. In *Proceedings of the 54th Hawaii International Conference on System Sciences (Proceedings of the Annual Hawaii International Conference on System*

Sciences), Tung Bui (Ed.). Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.739

[10] Joshua Gelhaar and Boris Otto. 2020. Challenges in the Emergence of Data Ecosystems. (2020).

[11] Andreas Hutterer and Barbara Krumay. 2022. Integrating Heterogeneous Data in Dataspaces - A Systematic Mapping Study. *PACIS 2022 Proceedings*. (2022).

[12] Maria Linnartz, Soo-Yon Kim, Martin Perau, Tobias Schröer, Sandra Geisler, and Stefan Decker. 2022. Unternehmensübergreifendes Datenqualitätsmanagement. *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 117, 12 (2022), 851–855. https://doi.org/10.1515/zwf-2022-1167

[13] Martin Lnenicka, Anastasija Nikiforova, Mariusz Luterek, Petar Milic, Daniel Rudmark, Sebastian Neumaier, Karlo Kević, Anneke Zuiderwijk, and Manuel Pedro Rodríguez Bolívar. 2024. Understanding the development of public data ecosystems: From a conceptual model to a six-generation model of the evolution of public data ecosystems. *Telematics and Informatics* 94 (2024), 102190. https://doi.org/10.1016/j.tele.2024.102190

[14] Marcelo Iury S. Oliveira and Bernadette Farias Lóscio. 2018. What is a data ecosystem?. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, Marijn Janssen, Soon Ae Chun, Vishanth Weerakkody, Anneke Zuiderwijk, and Charles C. Hinnant (Eds.). ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/3209281.3209335

[15] Marcelo Iury S. Oliveira, Lairson Emanuel R. A. Oliveira, Marlos G. Ribeiro Batista, and Bernadette Farias Lóscio. 2018. Towards a meta-model for data ecosystems. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, Marijn Janssen, Soon Ae Chun, Vishanth Weerakkody, Anneke Zuiderwijk, and Charles C. Hinnant (Eds.). ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/3209281.3209333

[16] B. Otto, S. Steinbuss, A. Teuscher, and S. Lohmann. [n.d.]. IDS Reference Architecture Model. https://doi.org/10.5281/zenodo.5105529

[17] Jan Pennekamp, Roman Matzutt, Christopher Klinkmüller, Lennart Bader, Martin Serror, Eric Wagner, Sidra Malik, Maria Spiß, Jessica Rahn, Tan Gürpinar, Eduard Vlad, Sander J. J. Leemans, Salil S. Kanhere, Volker Stich, and Klaus Wehrle. 2024. An Interdisciplinary Survey on Information Flows in Supply Chains. *Comput. Surveys* 56, 2 (2024), 1–38. https://doi.org/10.1145/3606693

[18] Minna M. Rantanen, Sami Hyrynsalmi, and Sonja M. Hyrynsalmi. 2019. Towards Ethical Data Ecosystems: A Literature Study. In *2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 1–9. https://doi.org/10.1109/ICE.2019.8792599

[19] Marcelo Iury S. Oliveira, Glória de Fátima Barros Lima, and Bernadette Farias Lóscio. 2019. Investigations into Data Ecosystems: a systematic mapping study. *Knowledge and Information Systems* 61, 2 (2019), 589–630. https://doi.org/10.1007/s10115-018-1323-6

[20] Anna Maria Schleimer and Estelle Duparc. 2025. Designing Digital Infrastructures for Industrial Data Ecosystems—A Literature Review. In *Solutions and Technologies for Responsible Digitalization*, Daniel Beverungen, Christiane Lehrer, and Matthias Trier (Eds.). Lecture Notes in Information Systems and Organisation, Vol. 75. Springer Nature Switzerland, Cham, 275–291. https://doi.org/10.1007/978-3-031-80122-8{_}18

[21] Syed Iftikhar Hussain Shah, Vassilios Peristeras, and Ioannis Magnisalis. 2021. Government Big Data Ecosystem: Definitions, Types of Data, Actors, and Roles and the Impact in Public Administrations. *Journal of Data and Information Quality* 13, 2 (2021), 1–25. https://doi.org/10.1145/3425709

[22] Sascha Welten, Marius de Arruda Botelho Herr, Lars Hempel, David Hieber, Peter Placzek, Michael Graf, Sven Weber, Laurenz Neumann, Maximilian Jugl, Liam Tirpitz, Karl Kindermann, Sandra Geisler, Luiz Olavo Da Bonino Silva Santos, Stefan Decker, Nico Pfeifer, Oliver Kohlbacher, and Toralf Kirsten. 2024. A study on interoperability between two Personal Health Train infrastructures in leukodystrophy data analysis. *Scientific data* 11, 1 (2024), 663. https://doi.org/10.1038/s41597-024-03450-6