Challenges and Opportunities for Enabling the Next Generation of Cross-Domain Dataspaces

Rohit A. Deshmukh[1], Diego Collarana[1], Joshua Gelhaar[3], Johannes Theissen-Lipp[1,2], Christoph Lange[1,2], Benedikt T. Arnold[1,2], Edward Curry[4] and Stefan Decker[1,2]

[1]Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany
[2]RWTH Aachen University, Aachen, Germany
[3]Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany
[4]Insight Centre for Data Analytics, University of Galway, Galway, Ireland

Abstract

Dataspaces are regarded as a standardized solution for sharing data in a trusted way. However, providing and sharing high-quality data across dataspaces poses several scientific and technical challenges, opening new research avenues. Developing governance models and technologies for supporting cross-domain integration of data and services from the existing single-domain dataspaces represents a significant challenge. In this vision paper, we discuss the challenges for enabling next-generation dataspaces and propose an innovative approach that aims at developing a vision and identifying requirements and building blocks for next-generation dataspaces, followed by defining a roadmap and practical migration paths for the existing dataspaces towards the next-generation dataspaces.

Keywords

Cross-Domain Dataspaces, Interoperability, Semantic Interoperability, AI, Large Language Models

1. Introduction

Dataspaces are essential for enabling data sharing among participants in a sovereign, interoperable, and trustworthy manner. They have gained prominence in Europe with the European Data Governance Act as part of the European Data Strategy[1]. Dataspaces set the path to a digital single market where data can flow seamlessly and securely across the borders and sectors within the EU. In the last few years, due to the increasing need to enable secure and interoperable data-sharing infrastructures among industries in the EU, dataspaces have been getting increasing attention from the research community and the industry. However, the siloed work of several independent initiatives has resulted in various definitions of dataspaces [1]. To prevent fragmented and heterogeneous implementations of dataspaces, the Data Spaces

The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, May 26 – 27, 2024, Hersonissos, Greece

rohit.deshmukh@fit.fraunhofer.de (R. A. Deshmukh); diego.collarana.vargas@fit.fraunhofer.de (D. Collarana)
0000-0003-2885-7076 (R. A. Deshmukh); 0000-0002-2583-0778 (D. Collarana); 0000-0002-8358-5424 (J. Gelhaar); 0000-0002-2639-1949 (J. Theissen-Lipp); 0000-0001-9879-3827 (C. Lange); 0000-0001-8594-880X (B. T. Arnold); 0000-0001-8236-6433 (E. Curry); 0000-0001-6324-7164 (S. Decker)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

Support Centre (DSSC)[2] is working towards identifying common guidelines and building blocks to accelerate the development of dataspaces in Europe. While the DSSC's harmonization of dataspace concepts enhances interoperability, establishing compatibility in a cross-domain dataspace scenario remains an open question.

Dataspaces are currently being established in specific domains[3], such as Manufacturing, Mobility, Culture, and Healthcare. Interoperability among such domain-specific dataspaces must be facilitated to enable an effective sharing and reuse of data across domains that will help companies develop new innovative business models and revenue streams. Research is required to establish the foundations and technology for this interoperation. This paper identifies gaps in the existing dataspace efforts, presents the challenges for enabling cross-domain dataspaces and proposes a novel approach that advocates for a coherent strategy combining semantic interoperability, human-centricity, trust, data stewardship and service quality.

2. Current Generation of Dataspaces

Several research initiatives develop architectures and technologies for dataspaces. For instance, the International Data Spaces (IDS) initiative[4] has developed the IDS Reference Architecture Model and the IDS Information Model [2] to enhance interoperability and sovereignty in data sharing. While Gaia-X[5] also shares these objectives, it additionally covers sovereignty over data in cloud infrastructures. However, neither IDS nor Gaia-X implements Persistent Identifiers (PIDs) (**Pitfall 1**), which have been a key enabler for interoperability in and across research data infrastructures. Therefore, transferring and evaluating this idea in the context of B2B industrial data infrastructures, or dataspaces is essential for addressing link rot and identifier clashes [3].

Most of the initiatives, including IDS and Gaia-X, advocate the use of semantics in dataspaces [4]; however, its potential is not yet fully utilized in actual implementations (**Pitfall 2**). Experts report[6] that interoperability is often seen only on the metadata level and not on the actual data payload level. There are no industry-specific controlled vocabulary and knowledge graphs in place today. It is, therefore, clear that an effective use of semantics is necessary for the success of dataspaces. However, the Semantic Web is perceived as complex by developers and its practical adoption is usually challenging for them [5, 2] (**Pitfall 3**). Therefore, there is an urgent need for new technological solutions to reduce this complexity and make the use of semantics in dataspaces user-friendly.

The three pitfalls mentioned above are exacerbated by the emergence of cross-domain data ecosystems, i.e., by the various interaction and exchange relationships among cross-domain dataspace participants. Data ecosystems enable data reuse, the integration of data users and data providers, and thus the linking of data to innovative services [6, 7]. Enabling a federation of dataspace ecosystems [8] leads to additional requirements concerning interoperability of metadata, data, identities, access policies, trust among participants, and pricing and governance models, etc., that need further research.

[2] https://dssc.eu/

[3] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

[4] https://www.internationaldataspaces.org/

[5] https://gaia-x.eu/

[6] https://www.trusts-data.eu/data-spaces-semantic-interoperability/workshop-report-pictures-slides/

3. Challenges and Opportunities for Next-Generation Dataspaces

In the evolution of dataspaces, enabling cross-domain interoperation holds the potential to enable new use cases (Figure 1). This involves several challenges and opportunities:

PID Infrastructure: PIDs play a crucial role in addressing link rot and identifier clashes [3] and facilitating interoperability. However, existing dataspace infrastructure concepts lack several requirements. The current PID infrastructure is openly accessible and centralized, which is problematic for competitive industries. Centralized systems also pose single points of failure. Examples of existing PID systems include ORCID.org, handle.net, and DOIs.

Data Quality, Incentivization, and Governance Frameworks: In the existing prevailingly domain-centric dataspaces involving participants at different levels of technical maturity, obtaining high-quality data is a big challenge. It takes a lot of effort, time, and several iterations to get the data providers to provide high-quality data. Organizational rules govern data provision, and typically, data providers do not receive any incentives, although their data allow for value-added services. Therefore, designing an economic model that incentivizes high-quality data provision and guarantees fair returns for providers is crucial.

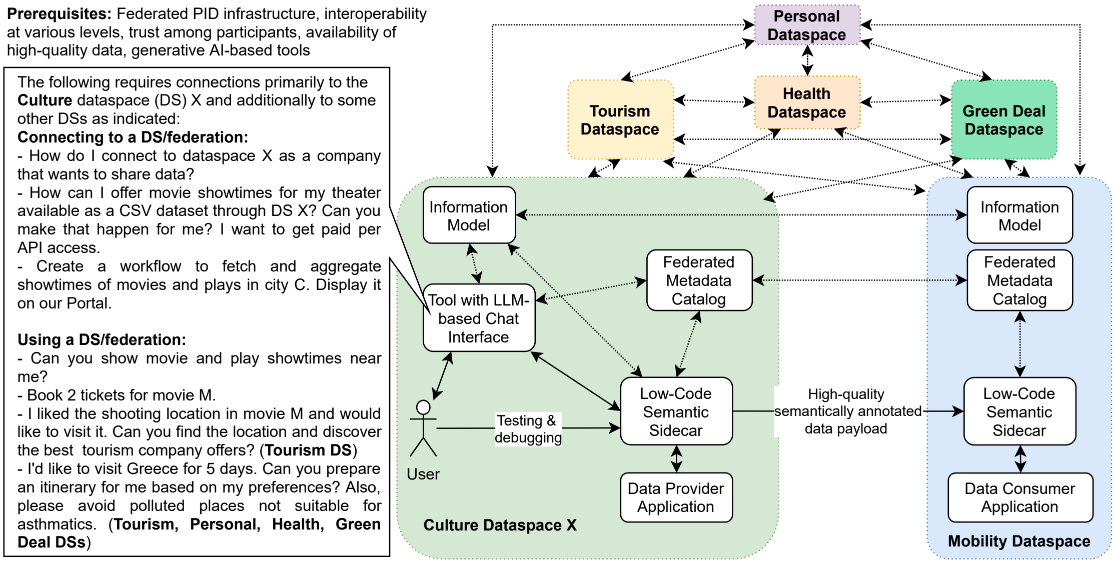
Sector-specificity vs Interoperability: Catering for the needs of sectors and to boost digitalization and innovation, it is necessary to enable and encourage bottom-up and sector-specific initiatives. While doing this, we must avoid silos, and ensure interoperability and seamless data flow across sectors and borders.

Complexity of Semantic Web and Practicality: Interconnecting dataspaces requires research on varying levels of granularity and technical depth (ranging from fine-grained service descriptions with Semantic Web standards to more abstract labelling frameworks [9]), for finding the balance between a practical integration practice and complexity of options. Early experiments with integrating data standards for domain-specific knowledge models [2] show that such an integration goes beyond an engineering challenge but poses scientific questions [5] regarding, e.g., how to delineate universal knowledge representation from application logic, and the scalability of declarative vs. procedural approaches.

The scientific challenge is to identify a logic for representing data semantics and communication protocols that is sufficiently expressive to capture relevant aspects of domain knowledge and sufficiently flexible to cope with the continuous evolution of data standards while remaining practically applicable for service providers given their scalability, compatibility, and compliance constraints. The technical challenge is to make essential services such as the persistent identification of artefacts scale across dataspaces and to define a process for migrating existing "live" dataspaces, into which participants have invested development efforts and where businesses are in operation, and to provide tooling support for executing this process with minimal disruption.

4. Approach for Enabling Next-Generation Dataspaces

Our approach to address technical, governance, and economic challenges consists of three steps: (1) Bottom-up investigation: A thorough exploration, analysis, and evaluation of existing dataspaces; (2) Top-down ideation: Development of a comprehensive vision for next-generation dataspaces in a green-field approach, i.e., not bound by limitations of legacy design and im-



Prerequisites: Federated PID infrastructure, interoperability at various levels, trust among participants, availability of high-quality data, generative AI-based tools

The following requires connections primarily to the **Culture** dataspace (DS) X and additionally to some other DSs as indicated:

Connecting to a DS/federation:

- How do I connect to dataspace X as a company that wants to share data?
- How can I offer movie showtimes for my theater available as a CSV dataset through DS X? Can you make that happen for me? I want to get paid per API access.
- Create a workflow to fetch and aggregate showtimes of movies and plays in city C. Display it on our Portal.

Using a DS/federation:

- Can you show movie and play showtimes near me?
- Book 2 tickets for movie M.
- I liked the shooting location in movie M and would like to visit it. Can you find the location and discover the best tourism company offers? (**Tourism DS**)
- I'd like to visit Greece for 5 days. Can you prepare an itinerary for me based on my preferences? Also, please avoid polluted places not suitable for asthmatics. (**Tourism, Personal, Health, Green Deal DSs**)

Personal Dataspace
Tourism Dataspace
Health Dataspace
Green Deal Dataspace
Information Model
Tool with LLM-based Chat Interface
Federated Metadata Catalog
Low-Code Semantic Sidecar
Testing & debugging
User
Data Provider Application
Culture Dataspace X
High-quality semantically annotated data payload
Information Model
Federated Metadata Catalog
Low-Code Semantic Sidecar
Data Consumer Application
Mobility Dataspace

Figure 1: Our vision for enabling the next generation of cross-domain dataspaces.

plementation decisions, and (3) Migration roadmap: Development of a strategic roadmap and practical migration paths from current dataspace systems towards next-generation dataspaces.

Addressing Complexity of Semantics in Dataspaces with User-friendly Tools and Interfaces: To foster technical interoperability, we propose adapting the widely accepted FAIR Data Principles (Findable, Accessible, Interoperable, Reusable) [10] originating from research data management to dataspaces and Knowledge Graphs [11]. Specifically, we aim to develop a distributed PID infrastructure for persistent identification and identifier mapping, including service descriptions with Semantic Web standards to more abstract "labeling frameworks". Furthermore, to bridge the gap between Semantic Web and dataspaces, our approach includes using tooling support. We plan to use intuitive, user-friendly tools to enable an effective use of semantics in dataspaces at every stage in the lifecycle of semantic data–from provision, semantic enrichment, and matchmaking to composition and consumption.

The latest advances in AI make it an excellent technology for improving user interface to dataspaces. We plan to explore using Large Language Models (LLMs) to automate tedious and repetitive tasks such as data mapping, translation, and semantic enrichment and integration, ultimately reducing costs and manual effort. Furthermore, the use of low-code development environments has been explored before to enable data integration and service composition, e.g., in the manufacturing domain [12]. We plan to enhance such proven tools with Semantic Web technologies and LLMs, particularly generative AI, to simplify complex and tedious operations in dataspaces. We also plan to investigate enhancing LLMs with Knowledge Graphs to reduce hallucinations, increase deterministic behavior and predictability and thus, data quality.

Addressing Data Quality with Incentivization: Based on the experiences gained from the existing dataspaces, we aim to establish a robust, flexible governance structure that facilitates seamless data interoperability, while protecting sovereignty, and ensuring fair data sharing.

Our governance model will also include the aspect of economics for incentivizing the provision of high-quality data. Firstly, we plan to develop a reward or incentive system that aligns with the value generated by data. This will encourage more parties to provide high-quality data. Secondly, we propose a mechanism to evaluate the true value of data by tracking its usage and assessing its value based on impact and usefulness. This mechanism could be complemented by trust value-assessment features to strengthen the built-in reward system further, ensuring fair compensation to data providers proportional to the value and trustworthiness of their data. Finally, we plan to explore cost recovery models for dataspace sustainability and new business models enabled by dataspaces, including data sharing, marketplaces, and value-added services, to create new revenue streams and monetize data innovatively.

Thus, our novel approach fosters a coherent strategy combining interoperability, user-friendliness, trust, incentivization, data stewardship and service quality. Future work includes implementing the presented approach, and developing and evaluating technological and process building blocks to ensure a smooth migration towards next-generation dataspaces.

Acknowledgments

Funded by EU Commission within 'Data Spaces Support Centre' (grant 101083412), Science Foundation Ireland (SFI) (grant SFI/12/RC/2289_P2) & German Federal Ministry of Education and Research (BMBF) within 'WestAI' (grant 01IS22094C) & 'FAIR Data Spaces' (grant FAIRDS05).

References

[1] E. Curry, S. Scerri, T. Tuikka, Data Spaces: Design, Deployment, and Future Directions, 2022.

[2] S. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A. W. Müller, H. Akyürek, M. Böckmann, B. T. Imbusch, J. Lipp, S. Geisler, C. Lange, The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content, in: ISWC, 2020.

[3] S. Auer, Semantic Integration and Interoperability, Springer Nature, 2022, pp. 195–210.

[4] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in Dataspaces: Origin and Future Directions, WWW '23 Companion, 2023.

[5] R. Verborgh, M. Vander Sande, The semantic web identity crisis: in search of the trivialities that never were, Semantic Web 11 (2020) 19–27.

[6] J. Gelhaar, T. Groß, B. Otto, A taxonomy for data ecosystems (2021).

[7] M. Jarke, B. Otto, S. Ram, Data sovereignty and data space ecosystems, 2019.

[8] B. Otto, The evolution of data spaces, in: Designing data spaces: The ecosystem approach to competitive advantage, Springer International Publishing Cham, 2022, pp. 3–15.

[9] Gaia-x policy rules and labelling document - 22.11 release., 2021. URL: https://docs.gaia-x.eu/policy-rules-committee/policy-rules-labelling/22.11/.

[10] M. Wilkinson, The FAIR guiding principle for scientific data management and stewardship (2016). URL: https://arrow.tudublin.ie/dataguide/2.

[11] P. A. Bonatti, S. Decker, A. Polleres, V. Presutti, Knowledge graphs: New directions for knowledge representation on the semantic web (Dagstuhl seminar 18371) (2019).

[12] R. A. Deshmukh, D. Jayakody, A. Schneider, V. Damjanovic-Behrendt, Data spine: A federated interoperability enabler for heterogeneous IoT platform ecosystems (2021).