

FAIRness in Dataspaces: The Role of Semantics for Data Management

Marco Hauff¹, Lina Molinas Comet², Paul Moosmann², Christoph Lange^{2,3},
Ioannis Chrysakis^{4,5,6} and Johannes Theissen-Lipp^{2,3}

¹Fraunhofer Institute for Integrated Circuits IIS, Nuremberg, Germany

²Fraunhofer Institute for Applied Information Technology FIT, Aachen, Germany

³RWTH Aachen University, Aachen, Germany

⁴IDLab, Department of Electronics and Information Systems, Ugent, imec, Belgium

⁵DTAI, Department of Computer Science, KU Leuven, Belgium

⁶Netcompany-Intrasoft, Research and Innovation Development Department, Luxembourg

Abstract

Effective data governance and management are necessary but challenging prerequisites for creating value from data assets. Findability, accessibility, interoperability, and reusability are guiding principles for data owners in managing and archiving datasets, known as the FAIR Principles. Dataspaces provide an infrastructure for heterogeneous, multi-source data integration and cross-organizational data sharing that would benefit from FAIR compliance. In this paper, we propose semantics as an approach to ensure data FAIRness, enabling machine-aided discovery and reuse of data in different formats and structures. We conduct a systematic literature review to translate the overarching principles into ten concrete methods that can be implemented using semantic technologies. In addition, we analyze three mature dataspace initiatives for their adherence to the FAIR Principles and describe their specific implementation. In summary, we argue that semantics provide a common and infrastructure-independent foundation for data management in emerging dataspaces.

Keywords

Dataspaces, Data Spaces, FAIR Data, Semantics, Data Sharing

1. Introduction

Data is a valuable strategic resource for competitiveness, driving innovation and the digital transformation of organizations. Business value is created by using and reusing data assets [1]. Consequently, organizations across all industries are adapting their strategies to incorporate data into their processes and take advantage of opportunities for optimization and automation [2]. Relevant data may originate from various sources belonging to different actors in the supply chain or from other market participants, which could also include competitors. Therefore,

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

✉ marco.hauff@iis.fraunhofer.de (M. Hauff); lina.teresa.molinas.comet@fit.fraunhofer.de (L. M. Comet); paul.moosmann@fit.fraunhofer.de (P. Moosmann); christoph.lange-bever@fit.fraunhofer.de (C. Lange); ioannis.chrysakis@ugent.be (I. Chrysakis); theissen-lipp@dbis.rwth-aachen.de (J. Theissen-Lipp)

🆔 0009-0006-1619-8762 (M. Hauff); 0000-0001-5446-6947 (L. M. Comet); 0009-0005-2114-8578 (P. Moosmann); 0000-0001-9879-3827 (C. Lange); 0000-0003-2665-4056 (I. Chrysakis); 0000-0002-2639-1949 (J. Theissen-Lipp)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

data-driven value is often created by combining datasets of multiple actors [3]. By facilitating cross-organizational data sharing, dataspace are becoming increasingly important for value co-creation from distributed data [4, 5]. Their potential is also being recognized at a political level, and governmental funding for dataspace initiatives such as Gaia-X encourages the development of secure infrastructures for data sharing [6].

Handling the increasing volume of unstructured data necessitates proper data governance and management practices. Typically, decisions on their concrete implementation are left to the data owner and remain undefined [7]. This heterogeneity poses a major challenge for dataspace. Therefore, various dataspace initiatives agree on the four principles of findability, accessibility, interoperability, and reuse (FAIR) as common data management practices to enable smooth integration of multi-source data [8]. However, the FAIR Principles are described at a general level and allow for various implementation options, complicating standardized data management. Semantics enrich data with context and thus enable automated search and processing. Using common, established semantic vocabularies can support FAIRness – i.e., compliance with the FAIR Data Principles – in dataspace [8, 9]. Previous literature deals with individual aspects of FAIR that need to be synthesized to provide a guideline for data management in dataspace. Therefore, we aim to answer the following research question: How can semantics contribute to FAIR compliance in dataspace?

The remainder of the paper is structured as follows. First, we detail the concept of dataspace, semantics, and the FAIR Principles. Second, we describe our methodological approach. Third, we present the semantic approach we identified for each FAIR Principle. Fourth, we assess three mature dataspace initiatives for their FAIR compliance. Finally, we conclude with a discussion of our findings.

2. Background

Dataspace. Database management systems are limited to structured data that correspond to a predefined schema. Consequently, these traditional systems are unsuitable for integrating and processing the increasing volume of heterogeneous data coming from multiple data sources [9, 10]. Franklin et al. [11] describe dataspace as an abstraction for managing data regardless of its format or structure. Unlike database management systems, dataspace do not force the complete integration of data. Instead, they adopt an incremental integration approach and gradually improve data accessibility and interoperability by leveraging Semantic Web technologies for metadata management [11, 12]. Recently, dataspace have been attributed immense potential for value co-creation from distributed data, providing a decentralized infrastructure for cross-organizational data sharing while maintaining data sovereignty [4, 13]. Large-scale, government-funded initiatives such as Gaia-X¹ aim to establish a standardized platform that facilitates data sharing and minimizes the effort required to integrate multi-source datasets. Repetitive and low-level administrative tasks such as search functionality, naming conventions, data lineage, and access management should be reduced and streamlined [9, 11]. So far, the dataspace initiatives are still in their infancy and standards have yet to be defined.

¹<https://gaia-x.eu>

FAIR Data Principles. Data governance and management are essential for the reusability of datasets. In many cases, individual procedures are applied that remain undefined and opaque, thus complicating or preventing data reuse by third parties. Wilkinson et al. [7] define four foundational principles, known as the FAIR data principles, to standardize inconsistent data management and archiving practices in the scientific environment. Within the FAIRification process, digital resources become manageable. The underlying four principles are interconnected but can be implemented independently of each other. Besides the actual data, the process also includes the algorithms, tools, and workflows relevant to data collection. Overall, compliance with the FAIR Principles reduces human intervention by enabling machines to automatically discover, process, and integrate data [7, 14]. Efficient data management in line with the FAIR Principles is considered particularly important for dataspace to facilitate data sharing [15].

Semantics. A common language for describing data assets is crucial for managing and sharing data in dataspace [10]. Semantics are predefined vocabularies to describe relationships between data entities in a machine-readable format, supporting data integration and enabling automated data discovery [16]. The variety of terminologies poses a critical challenge for communication and uniform understanding between different parties [10]. Seamless data integration from multiple sources requires semantic interoperability, i.e., a shared understanding of the vocabulary. Ontologies aim to reduce semantic heterogeneity through standardized and generally accepted vocabularies [17].

3. Methodology

This paper investigates and answers the research question through a systematic literature review. The process is based on the structures proposed by vom Brocke et. al [18] and Webster & Watson [19], which describe the path from general to specific results. Following vom Brocke et. al [18], we begin our research by conceptualizing the topic and delineating the search fields: ‘dataspace’, ‘FAIR Data Principles’, and ‘semantics’. The purpose of this study is to examine the correlation between semantics and FAIR dataspace. To achieve this, we consider each of the FAIR Data Principles separately. Thus, we use the search string: *"dataspace*" OR "data Space*" AND ("FAIR" OR "find*" OR "access*" OR "interoper*" OR "reuse*") AND "semantic"*. In the second step, we search the following databases for relevant literature: ACM Digital Library, AIS eLibrary, IEEE Xplore, and Scopus. For ACM and Scopus, we limit the search to the title, keywords, and abstract. The search was conducted in February 2024. ACM yielded n=30 articles, AIS n=58, IEEE Xplore n=225, and Scopus n=163 articles. We evaluate the identified articles in an initial review by scanning the title, abstract, and keywords for relevance. Articles that do not address any of our defined search fields are excluded. Duplicates and non-peer-reviewed articles are also removed. After this first phase, we are left with n=78 articles. We then perform an in-depth screening of the abstract, methodology, results, discussion, and conclusion, removing those that do not address synergies between semantics and FAIR dataspace. After this screening, n=26 articles remain. In addition, we perform a forward and backward search as proposed in [19], resulting in n=21 additional articles. These articles are analyzed in the same cycle as those retrieved directly from the search string, resulting in a total of n=37 articles.

4. Results

From the literature, we extract the semantic approaches that are applicable in dataspace. We code and list the ten most frequently mentioned ones, assign them to the underlying FAIR Principle according to their intended use and create from this the requirements profile for dataspace, as shown in Table 1. Then, based on the analysis of architecture documents and code repositories, we compare the extent to which the semantics in the three currently most popular dataspace initiatives ensure FAIRness according to these criteria.

4.1. FAIRness Requirements

Findability. To ensure findability in dataspace, three requirements can be met by semantics: *self-descriptions*, *metadata*, and *catalogs*. Digital resources require globally unique and persistent identifiers (PIDs), such as Uniform Resource Identifiers, specified in RFC 3986². These identifiers are widely used in current initiatives, such as Gaia-X, allowing resources to be addressed, modified, and shared separately, even across multiple dataspace [8]. In addition to PIDs, the *self-description* of resources also includes information about the entities' attributes and relationships, often using subject-predicate-object triples to comply with *metadata* [20, 21]. Therefore, the Resource Description Framework (RDF) data model is commonly used to model these data structures. Using RDF vocabularies, such as the Data Catalog Vocabulary (DCAT)³, dataspace participants can exchange standardized information, including asset name, type, and lineage. This results in data structures that are both human and machine actionable. *Catalogs* and search tools, such as the RDF query language SPARQL, allow users to access shared information and browse descriptions to find valuable data for their specific use case [22].

Accessibility. Semantics are crucial for accessibility in FAIR-compliant dataspace, particularly regarding *authentication*, *authorization*, and *pipeline*. Like assets, dataspace participants receive a unique and descriptive identity, managed and verified by external identity providers. Trust between identities is ensured through cryptographic verification, which allows only *authenticated* actors to engage within the dataspace [23, 24]. Due to the sensitivity of data, global sharing is not desired. *Authorization* for access and usage rights is required. Thus, participants manage their offerings, data or services, via policy frameworks to determine the use and scope of data. Standards such as the Open Digital Rights Language (ODRL) facilitate the creation, communication, and execution of contracts in the data space [25, 26]. Leveraging semantics enables machines to understand interfaces and processes, and thus allows for *automation* [27, 28].

Interoperability. Dataspace promote the exchange of data between parties. To ensure a seamless process, dataspace must be interoperable. This requires participants to exchange data to be able to comprehend and integrate them. Semantics can aid in *standardization* and *integration*. Usually, dataspace offer a vocabulary hub, providing an overview of commonly established and *standardized* terms and vocabularies as examples and best practices for describing data, services, and contracts [37, 46]. The hub facilitates the exchange of documentation among

²<https://datatracker.ietf.org/doc/html/rfc3986>

³<https://www.w3.org/TR/vocab-dcat-3/>

Table 1
Requirements for FAIRness in dataspace

Requirement	Objective	References
FINDABLE		
Self-Description	Use unique identifiers and metadata self-descriptions to accurately and unambiguously characterize participants, services, and assets.	8, 20, 21, 22, 29
Metadata	Use standard vocabularies and ontologies, to ensure data is described accurately and comprehensibly.	10, 21, 29, 30, 31, 32, 33
Catalog	Use a publicly available catalog that aggregates participants, services and assets to enable discoverability.	11, 21, 34, 35, 36, 37, 38, 39
ACCESSIBLE		
Authentication	Use a self-description to participate in the dataspace that can be verified by an external identity provider	23, 24, 29, 40, 41
Authorization	Use policies to manage data sharing among participants and determine the constitutions and parties involved.	25, 26, 29, 42, 43, 44
Pipeline	Use service descriptions to automate interactions and processing pipeline.	20, 21, 27, 28, 45
INTEROPERABLE		
Standardization	Use shared ontologies and community standards to ensure consistent understanding and interoperability.	10, 29, 30, 32, 33, 37, 46, 47
Integration	Use reasoning and aggregation techniques to incorporate heterogeneous data sources.	10, 36, 48, 49, 50, 51
REUSABLE		
Reliability	Use verification and validation tools to enhance the integrity, quality, and usability of data assets.	32, 48, 52, 53, 54.
Enrichment	Use contextual metadata annotations to enhance data assets and facilitate advanced queries and links.	26, 29, 31, 36, 45, 48, 52

relevant parties, promoting a shared understanding [29]. Therefore, it enables the *integration* of data from different sources, such as multiple datasets in a structured format (e.g., RDF), which can be aggregated and queried together. This allows to reveal previously unknown connections between data [50, 51].

Reusability. All of the previously mentioned principles contribute to the concept of reusability. For example, if an entity is easily findable, it is more likely to be reused. Similarly, easily comprehensible entities by following a common standard increase the likelihood of reuse. Additionally, semantics can improve the *reliability* and *enrichment* of data. *Reliability* assessment and assurance can be achieved by using shapes to verify the format of incoming data and automating

standardization, as mentioned under interoperability. RDF validation is often performed against shapes defined in the Shapes Constraint Language (SHACL) or Shape Expressions (ShEx), which define the requirements that an RDF graph must meet [53, 54]. *Enrichment* can be applied throughout the data lifecycle, mapping raw data into a structured format and augmenting it with additional information from external sources [45].

4.2. State of Development

Several endeavors aim to define principles and guidelines for the creation and development of dataspace and to transfer concepts from theory into practice. Today, several well-known initiatives in various application areas have reached distinct stages of development and maturity. Initially, such initiatives define the requirements, principles and specifications to be considered when implementing dataspace. Subsequently, these specifications can lead to implementations or elements that facilitate the establishment of dataspace in which different actors or participants can exchange data. These implementations can then be sector-specific, such as Catena-X⁴ in the automotive sector, in Industry 4.0⁵ with a focus on manufacturing, Prometheus-X⁶ in the education and skill sector, or with more general developments such as FAIR Data Spaces⁷ at the interface between research and various application domains.

It is also worth mentioning that additional projects looking to facilitate the implementation of dataspace exist, such as the Eclipse Dataspace Components (EDC)⁸, whose adoption in several projects is documented. The EDC is a framework for data sharing in a cross-organizational and sovereign manner supporting specifications from IDS and Gaia-X. One of its main components is the EDC Connector, a well-defined interface exposing the dataspace participants' back-office infrastructure, aiming to bring together their otherwise incompatible and without access/usage control infrastructure. The connector provides functionalities such as discovering, connecting, automated contract negotiation, policy enforcement, and auditing processes. The Minimum Viable Dataspace⁹ is a sample implementation of the EDC, leveraging it and showing its capabilities. Currently, these two do not support semantics, but one could extend the connector to include such functionality. Other implementations of connectors exist, such as those listed in [55]. The Sovity Connector¹⁰, based on the EDC and the Dataspace Connector, extends the functionality of the EDC. For example, it allows the EDC to communicate with the catalog called *IDS Metadata Broker*¹¹.

In this paper, we solely focus on assessing the leading and most frequently mentioned initiatives in research. The most commonly noted initiatives in research include Gaia-X, International Data Spaces (IDS), and the European Open Science Cloud (EOSC). These initiatives define the architectural frameworks for dataspace [56]. Our assessment of the initiatives consists of reviewing the official documentation of the specifications and the repositories with available

⁴<https://catena-x.net/en/about-us/operating-environment-1>

⁵<https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/PositionPaper-DataSpace.html>

⁶<https://dataspace.prometheus-x.org/>

⁷<https://www.nfdi.de/fair-data-spaces/?lang=en>

⁸<https://github.com/eclipse-edc>

⁹<https://github.com/eclipse-edc/MinimumViableDataspace>

¹⁰<https://github.com/sovity>

¹¹<https://github.com/International-Data-Spaces-Association/metadata-broker-open-core>

code. In this way, we want to find out whether the written specifications reflect concrete implementations and how these align with the dataspace requirements introduced in Table 1. The results of our assessment are shown in Table 2. There, we summarize our findings as follows: the symbol ✓ indicates that the requirement is fully implemented, (✓) shows that the requirement is only partially supported, i.e., it is only present as a specification but not yet implemented. However, a * indicates that the requirement is not even part of the specification.

Table 2

The evaluated FAIRness of current dataspace initiatives

	Gaia-X	IDS	EOSC
Self-Description	✓	✓	(✓)
Metadata	(✓)	✓	✓
Catalog	✓	✓	✓
Authentication	✓	✓	(✓)
Authorization	✓	✓	(✓)
Pipeline	*	*	*
Standardization	✓	✓	✓
Integration	(✓)	*	*
Reliability	✓	✓	*
Enrichment	*	*	✓

Gaia-X. Gaia-X aims to establish an ecosystem, whereby data is shared and made available in a trustworthy environment. To achieve this goal, Gaia-X defines several federation services, where a federation implements and federates a dataspace. These services are grouped into four sets, namely *Identity & Trust*, *Federated Catalogue*, *Sovereign Data Exchange*, and *Compliance*¹². One major reference implementation of federation services is provided by the XFSC (Cross Federation Service Components) repositories under the Eclipse Foundation¹³. The *Identity & Trust* set of services contains an Authentication/Authorization service, which implements these requirements as defined in Table 1. The *Federated Catalogue* set consists of the Federated Catalogue as well as tool support for self-descriptions. The implementation of the XFSC Federated Catalogue enables the management of self-descriptions and also allows for the validation against SHACL shapes as a measure towards reliability. The topic of Metadata and Standardization are addressed in the Gaia-X Trust Framework¹⁴. In this framework, a Gaia-X ontology is specified, which has to be used to describe all participants and services of a Gaia-X dataspace. Additionally, constraints are specified, which are modelled using SHACL. All ontology and SHACL graphs can be accessed using the Gaia-X Registry¹⁵. The Gaia-X specifications also specify the re-use of certain metadata vocabularies such as DCAT, for example, for the data exchange services¹⁶. The topic of integration is partly specified by the notion of service

¹²<https://www.gxfs.eu/set-of-services/>

¹³<https://gitlab.eclipse.org/eclipse/xfsc/>.

¹⁴<https://docs.gaia-x.eu/policy-rules-committee/trust-framework/22.10/>

¹⁵<https://registry.lab.gaia-x.eu/v1/docs>

¹⁶<https://docs.gaia-x.eu/technical-committee/data-exchange/latest/dewg/>

compositions¹⁷, which allow the aggregation of several services. This way, a service can re-use data that emerged from the application of another service to apply further processing steps. The remaining requirements defined in Table 1, namely pipeline and enrichment, are currently not addressed by Gaia-X specifications or implementations.

IDS. The latest developments of the IDS Dataspace Protocol¹⁸ reflect the specifications regarding the Dataspace Information Model, which covers the definitions of the main concepts to be considered in IDS-based dataspaces. The implementation of connectors in an IDS dataspace context must respond with a JSON-LD data object compliant with JSON Schemas and SHACL shapes. Moreover, participants describe themselves and their resources and infrastructure. These self-descriptions can be registered, published, queried, and maintained by the IDS Metadata Broker. The self-descriptions are metadata, and the use of existing semantic web standards is favored. In this sense, the IDS Information Model [23] is modeled as an RDF/OWL ontology and reuses concepts of the DCAT, ODRL, Time, DQV, and other vocabularies¹⁹. Moreover, SHACL shapes are available for testing²⁰. There is also an implementation of the Dataspace Connector²¹, which uses the IDS Messaging Services for the functionalities and message handling, as specified in the IDS Reference Architecture Model 4.0²² and integrates the IDS Information Model. Moreover, the Catalog Protocol specifies how a data consumer requests a catalog from a catalog service. Such a catalog is DCAT and ODRL compliant. In the latest version of IDS RAM 4.0, the Identity Authority role includes specifications about the authorization functionalities and the clearing house²³, which serves as an intermediary to provide clearing and settling services for the data exchange transactions in the IDS. Such a component is similar to the Data Exchange Logging Service from Gaia-X²⁴. Additionally, the Dynamic Attribute Provisioning Service²⁵ is part of the Identity Provider to verify the attributes of the participants and connectors in the dataspace.

EOSC. The EOSC initiative addresses the FAIR Principles, with interoperability as a core concept, and aims to create a shared data space for research, science and innovation data management while ensuring the protection of data through EU laws [57]. The specified requirements regarding semantic interoperability are the following: there should be a definition of the concepts, their metadata and data schemas, and they should be publicly available; semantic artefacts should be FAIR, available preferably using open licenses, and have associated documentation, and support maintenance; a metadata model, based on existing standards, should be available to allow discovery over existing federated research data and metadata; there should be building blocks and protocols to facilitate the federation and harvesting of semantic artefacts catalogs.

¹⁷docs.gaia-x.eu/technical-committee/architecture-document/latest/component_details/#service-composition

¹⁸<https://docs.internationaldataspaces.org/ids-knowledgebase/v/dataspace-protocol/overview/readme>

¹⁹<https://international-data-spaces-association.github.io/InformationModel/docs/index.html>

²⁰<https://github.com/International-Data-Spaces-Association/InformationModel/tree/develop/testing>

²¹<https://github.com/International-Data-Spaces-Association/DataspaceConnector>

²²https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3-1-business-layer/3_1_1_roles_in_the_ids

²³<https://github.com/International-Data-Spaces-Association/IDS-G/tree/main/Components/ClearingHouse>

²⁴http://docs.gaia-x.eu/technical-committee/architecture-document/latest/enabling_services/

²⁵<https://github.com/International-Data-Spaces-Association/IDS-G>

The technical layer defines a common security and privacy framework covering authorization and authentication functionalities. The FAIRCORE4EOSC²⁶ project focuses on the development of core components²⁷ for the EOSC namely the Compliance Assessment Toolkit to provide services related to policies and vocabulary services; the EOSC Data Type Registry to register the PID metadata elements including provenance information; the Metadata Schema and Crosswalk Registry to allow register users to create, register and version schemas and crosswalks with PIDs; the EOSC PID Meta Resolver to map items into records; the Research Activity Identifier Service to provide persistent identifiers for research projects; the EOSC Research Discovery Graph Service to allow discovery of EOSC elements from the catalog (resources and communities); the EOSC Research Software APIs and Connectors to guarantee the enduring preservation of research software across various disciplines. Although metadata registration and vocabularies services are mentioned, the concept of self-descriptions is not noted.

Summary. Table 2 shows that two of the initiatives, Gaia-X and IDS, satisfy most of the requirements for semantics, while EOSC has specifications for such requirements, but they are not all implemented. It also shows that these initiatives mostly lack specifications for the more specific functionalities such as *pipeline*, *integration*, and *enrichment* requirements for dataspace, except EOSC which offers more advanced features covering research discovery through their metadata. Such findings indicate that the initiatives are currently focused on defining the main elements of dataspace. These efforts cover aspects like authentication and authorization, offering a catalog of available resources for transfer, ensuring self-descriptions and metadata to identify these resources, promoting standardization by using existing (W3C) standards, and providing SHACL shapes for validation and reliability.

Another important remark, based on our research of the repositories of the initiatives and connectors, is that some IDS components, such as the IDS Connector are not currently maintained in their original repositories. However, Sovity is supporting the maintenance of the IDS Connector, and extending some of its functionalities. Regarding Gaia-X, some specifications and implementations have not been updated since their initial release, e.g., the implementations of the XFSC components implement the specifications given by the 21.03 version of the Trust Framework, which has since then been replaced by the newer 22.10 version²⁸.

²⁶<https://faircore4eosc.eu/>

²⁷<https://faircore4eosc.eu/eosc-core-components>

²⁸<https://docs.gaia-x.eu/framework/?tab=software>

5. Discussion and Conclusion

In our paper, we derive a framework to assess the semantic FAIRness of dataspace, which we directly apply to investigate the FAIRness of three mature initiatives. Our framework provides a basis for rating further initiatives and third-party developments that also deal with the development of dataspace. Research and innovation projects such as MobiSpaces [58], Green.Dat.AI²⁹ or Flex4Res³⁰, which aim to use dataspace technology, have not yet been assessed. The same applies to projects funded by companies or through private programs that already provide the essential elements for setting up dataspace, e.g., Solid [59].

Theoretical Contribution. This paper synthesizes the existing literature to provide a comprehensive overview of the evolution and current state of semantics in dataspace. By extending the FAIR Principles, we propose a framework with tangible requirements that provide semantic clarity and interoperability. Accordingly, our paper establishes a foundation for evaluating future dataspace approaches based on the extended FAIR Principles, promoting a structured and objective assessment process.

Practical Implication. Connector implementations can leverage our extended FAIR framework to derive the requirements that a dataspace should fulfill. Vocabularies could be further developed and concretized specifically for the requirements found to cover potential demands.

Limitations. The analysis is limited to a specific number of dataspace developments, potentially missing emerging trends and niche innovations. Insight into the development and operational intricacies of these dataspace is limited, relying on publicly available information and academic publications. Standardization processes within dataspace often have a pay-as-you-go approach, which introduces variability in the implementation and adherence to the proposed FAIR Principles, affecting the generalizability of our findings.

Conclusion. Our literature review suggests that semantic approaches in theory and practice can guide the FAIRness of dataspace. Semantic tools focus on standardization, establishing a uniform understanding through shared and standardized vocabularies. They support all FAIR Data principles and contribute to improvements. Utilizing standardized identifiers, shared ontologies, and rich contextual metadata, semantics enhance the individual aspects of the FAIR Principles and create a more interconnected and efficient data ecosystem. Current dataspace initiatives demonstrate similar approaches, with most FAIR Principles already guaranteed by semantics. However, there are additional approaches available, especially in automation such as our identified requirements pipeline, integration, and enrichment. Future directions could see dataspace evolving to facilitate automatic data exchange and analysis, thereby improving the efficiency of data use. Capable and reliable integration systems can provide a larger data basis that can be further enhanced through enrichment. This includes further research into the development of advanced, dynamic ontologies and the automation of ontology matching

²⁹<https://greendataai.eu/>

³⁰<https://www.flex4res.eu/>

for the seamless integration of different data sources. To ensure scalability, the continuous optimization of graph databases and the use of parallel processing to efficiently manage large amounts of data are essential. In addition, developing secure, ethical semantic technologies to protect privacy and promote responsible data use is a necessity if stakeholders are to share their data.

Acknowledgments

This publication is based upon work from COST Action DKG (CA19134), supported by COST (European Cooperation in Science and Technology). This work has been partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the Antrieb 4.0 project (Grant No. 13IK015B), by the European Union's funded Projects MobiSpaces (Grant agreement no 101070279), Green.DAT.AI (Grant agreement no 101070416), AgriDataValue (Grant agreement no 101086461) and by the FAIR Data Spaces project of the German Federal Ministry of Education and Research (BMBF) under the grant number FAIRDS05.

References

- [1] F. von Scherenberg, M. Hellmeier, B. Otto, Data sovereignty in information systems, *Electronic Markets* 34 (2024) 1–11. doi:10.1007/s12525-024-00693-4.
- [2] M. Rüßmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel, M. Harnisch, *Industry 4.0: The future of productivity and growth in manufacturing industries*, Boston consulting group 9 (2015) 54–89.
- [3] J. Gelhaar, B. Otto, Challenges in the emergence of data ecosystems, in: *PACIS 2020 Proceedings*, 2020. URL: <https://aisel.aisnet.org/pacis2020/175>.
- [4] C. Cappiello, A. Gal, M. Jarke, J. Rehof, Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391), *Dagstuhl Reports* 9 (2020) 66–134. doi:10.4230/DagRep.9.9.66.
- [5] P. Singh, M. J. Beliatas, M. Presser, Enabling edge-driven dataspace integration through convergence of distributed technologies, *Internet of Things* 25 (2024) 1–33. doi:10.1016/j.iot.2024.101087.
- [6] A. Seidel, K. Wenzel, A. Hänel, ..., H. Ernst, Towards a seamless data cycle for space components: considerations from the growing european future digital ecosystem gaia-x, *CEAS Space Journal* (2023) 1–14. doi:10.1007/s12567-023-00500-4.
- [7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, ..., B. Mons, The fair guiding principles for scientific data management and stewardship, *Scientific Data* 3 (2016) 1–9. doi:10.1038/sdata.2016.18.
- [8] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in dataspace: Origin and future directions, in: *Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion*, Association for Computing Machinery, New York, NY, USA, 2023, p. 1504–1507. doi:10.1145/3543873.3587689.
- [9] E. Curry, Future research directions for dataspace, data ecosystems, and intelligent

systems, Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems (2020) 297–304. doi:10.1007/978-3-030-29665-0_18.

- [10] A. Halevy, A. Rajaraman, J. Ordille, Data integration: the teenage years, in: Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB '06, VLDB Endowment, 2006.
- [11] M. Franklin, A. Halevy, D. Maier, From databases to dataspace: a new abstraction for information management, *ACM Sigmod Record* 34 (2005) 27–33. doi:10.1145/1107499.1107502.
- [12] A. Halevy, Z. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov, The piazza peer data management system, *IEEE Transactions on Knowledge and Data Engineering* 16 (2004) 787–798. doi:10.1109/TKDE.2004.1318562.
- [13] N. C. Kuicheu, N. Wang, F. T. G. Narcisse, D. Xu, S. Francois, Building semantic relationships incrementally in dataspace, in: 2009 First International Conference on Information Science and Engineering, IEEE, 2009. doi:10.1109/ICISE.2009.370.
- [14] M. Scheffler, M. Aeschlimann, M. Albrecht, ..., C. Draxl, Fair data enabling new horizons for materials research, *Nature* 604 (2022) 635–642. doi:10.1038/s41586-022-04501-x.
- [15] A. Kotsev, M. Minghini, R. Tomas, V. Cetl, M. Lutz, From spatial data infrastructures to data spaces—a technological perspective on the evolution of european sdIs, *ISPRS International Journal of Geo-Information* 9 (2020) 1–19. doi:10.3390/ijgi9030176.
- [16] E. Curry, S. Scerri, T. Tuikka, Data Spaces: Design, Deployment, and Future Directions, Springer International Publishing, Cham, 2022. doi:10.1007/978-3-030-98636-0_1.
- [17] K.-D. Schewe, B. Thalheim, Semantics in data and knowledge bases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-88594-8_1.
- [18] J. v. Brocke, A. Simons, B. Niehaves, B. Niehaves, K. Reimer, R. Plattfaut, A. Clevén, Reconstructing the giant: On the importance of rigour in documenting the literature search process, in: ECIS 2009 Proceedings, European Conference on Information Systems, 2009. URL: <https://aisel.aisnet.org/ecis2009/161>.
- [19] J. Webster, R. T. Watson, Analyzing the past to prepare for the future: Writing a literature review, *MIS Quarterly* 26 (2002) xiii–xxiii. URL: <http://www.jstor.org/stable/4132319>.
- [20] S. P. Stier, X. Xu, L. Gold, M. Möckel, Ontology-based battery production dataspace and its interweaving with artificial intelligence-empowered data analytics, *Energy Technology* (2024) 8–13. doi:10.1002/ente.202301305.
- [21] H. B. Nasrabadi, T. Hanke, M. Weber, ..., B. Skrotzki, Toward a digital materials mechanical testing lab, *Computers in Industry* 153 (2023) 1–15. doi:10.1016/j.compind.2023.104016.
- [22] O. Hartig, C. Bizer, J.-C. Freytag, Executing sparql queries over the web of linked data, in: The Semantic Web - ISWC, 2009. doi:10.1007/978-3-642-04930-9_19.
- [23] S. Bader, J. Pullmann, C. Mader, ..., C. Lange, The international data spaces information model – an ontology for sovereign exchange of digital content, *Lecture Notes in Computer Science* 12507 LNCS (2020) 176 – 192. doi:10.1007/978-3-030-62466-8_12.
- [24] S. Stubblebine, R. Wright, An authentication logic with formal semantics supporting synchronization, revocation, and recency, *IEEE Transactions on Software Engineering* 28 (2002) 256–285. doi:10.1109/32.991320.
- [25] H. J. M. Bastiaansen, S. Dalmolen, M. Kollenstart, T. M. van Engers, User-centric network-

- model for data control with interoperable legal data sharing artefacts, in: Pacific Asia Conference on Information Systems, 2020. doi:10.1016/j.compind.2023.104016.
- [26] S. Opriel, F. Möller, U. Burkhardt, B. Otto, Requirements for usage control based exchange of sensitive data in automotive supply chains, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021. doi:10.24251/HICSS.2021.051.
 - [27] F. Amato, V. Casola, A. Gaglione, A. Mazzeo, A semantic enriched data model for sensor network interoperability, *Simulation Modelling Practice and Theory* 19 (2011) 1745–1757. doi:10.1016/j.simpat.2010.09.010.
 - [28] F. Burzlaff, C. Bartelt, A conceptual architecture for enabling future self-adaptive service systems, in: 52nd Hawaii International Conference on System Sciences, HICSS 2019, Atlanta, GA, 2019. URL: <https://madoc.bib.uni-mannheim.de/49901/>.
 - [29] C. Schwede, J. Cirullies, On-demand shared digital twins – an information architectural model to create transparency in collaborative supply networks, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, pp. 1675–1684. doi:10.24251/HICSS.2021.202.
 - [30] C. Meghini, Linked open data & metadata, in: Handbook of Digital Public History, De Gruyter Oldenbourg, Berlin, Boston, 2022. doi:doi:10.1515/9783110430295-039.
 - [31] D. Paparova, Exploring the ontological status of data: A process-oriented approach, in: ECIS 2023 Research Papers, 2023. URL: https://aisel.aisnet.org/ecis2023_rp/299.
 - [32] S. Scheider, F. Lauf, F. Möller, B. Otto, A reference system architecture with data sovereignty for human-centric data ecosystems, *Business & Information Systems Engineering* 65 (2023) 577–595. doi:10.1007/s12599-023-00816-9.
 - [33] T. Wessel, K. Heuing, M. Schlangen, B. Schnieders, M. Algermissen, Rare diseases, digitization, and the national action league for people with rare diseases (namse), *Bundesgesundheitsblatt* 65 (2022) 1119 – 1125. doi:10.1007/s00103-022-03597-w.
 - [34] E. Curry, E. Curry, Fundamentals of real-time linked dataspace, *Real-time Linked Dataspace: Enabling Data Ecosystems for Intelligent Systems* (2020) 63–80.
 - [35] N. Jahnke, B. Otto, Data catalogs in the enterprise: Applications and integration, *Datenbank-Spektrum* 23 (2023) 89–96. doi:10.1007/s13222-023-00445-2.
 - [36] M. Franklin, A. Halevy, D. Maier, A first tutorial on dataspace, *Proc. VLDB Endow.* 1 (2008) 1516–1517. doi:10.14778/1454159.1454217.
 - [37] J. Möller, D. Jankowski, A. Hahn, Towards an architecture to support data access in research data spaces, in: 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), 2021. doi:10.1109/IRI51335.2021.00049.
 - [38] J. Umbrich, M. Karnstedt, J. X. Parreira, A. Polleres, M. Hauswirth, Linked data and live querying for enabling support platforms for web dataspace, in: 2012 IEEE 28th International Conference on Data Engineering Workshops, 2012. doi:10.1109/ICDEW.2012.55.
 - [39] S. V. Manisekaran, J. Sathishkumar, A fuzzy based semantic search engine for document retrieval in a personalized data space, in: International Conference on Recent Trends in Information Technology (ICRTIT), 2016. doi:10.1109/ICRTIT.2016.7569577.
 - [40] Á. Alonso, A. Pozo, J. M. Cantera, F. De la Vega, J. J. Hierro, Industrial data space architecture implementation using fiware, *Sensors* 18 (2018) 1–18. doi:10.3390/s18072226.
 - [41] I. Elsayed, P. Brezany, A. Tjoa, Towards realization of dataspace, in: 17th International

Workshop on Database and Expert Systems Applications (DEXA'06), 2006. doi:10.1109/DEXA.2006.140.

- [42] J. Hernandez, L. McKenna, R. Brennan, Tikd: A trusted integrated knowledge dataspace for sensitive healthcare data sharing, in: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021. doi:10.1109/COMPSAC51774.2021.00280.
- [43] L. Jin, Y. Zhang, X. Ye, An extensible data model with security support for dataspace management, in: 2008 10th IEEE International Conference on High Performance Computing and Communications, 2008. doi:10.1109/HPCC.2008.70.
- [44] P. de Alencar Silva, R. Fadaie, M. van Sinderen, Towards a digital twin for simulation of organizational and semantic interoperability in ids ecosystems, Proceedings of the Workshop of I-ESA 3214 (2022). URL: <https://api.semanticscholar.org/CorpusID:252599898>.
- [45] L. Sánchez, J. Lanza, J. R. Santana, ..., N. Crespi, Data enrichment toolchain: A data linking and enrichment platform for heterogeneous data, IEEE Access 11 (2023) 103079–103091. doi:10.1109/ACCESS.2023.3317705.
- [46] C. Roda, E. Navarro, C. E. Cuesta, A comparative analysis of linked data tools to support architectural knowledge, in: Integrated Spatial Databases, 2014. URL: <https://api.semanticscholar.org/CorpusID:10317475>.
- [47] W. Lin, C. Hu, Y. Li, X. Cheng, Virtual dataspace – a service oriented model for scientific big data, in: 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies, 2013. doi:10.1109/EIDWT.2013.5.
- [48] S. R. Jeffery, M. J. Franklin, A. Y. Halevy, Pay-as-you-go user feedback for dataspace systems, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, Association for Computing Machinery, 2008, pp. 847–860. doi:10.1145/1376616.1376701.
- [49] Y. Li, X. Meng, Research on personal dataspace management, in: Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research, IDAR '08, Association for Computing Machinery, 2008. doi:10.1145/1410308.1410311.
- [50] H. Belani, P. Šolić, T. Perković, An industrial iot-based ontology development for well-being, aging and health: A scoping review, in: 2022 IEEE International Conference on E-health Networking, Application & Services (HealthCom), 2022. doi:10.1109/HealthCom54947.2022.9982769.
- [51] N. Dessi, B. Pes, Towards scientific dataspaces, in: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, volume 3, 2009. doi:10.1109/WI-IAT.2009.353.
- [52] R. A. Buchmann, D. Karagiannis, Enriching linked data with semantics from domain-specific diagrammatic models, Business & Information Systems Engineering 58 (2016) 341–353. doi:10.1007/s12599-016-0445-1.
- [53] S. Staworko, I. Boneva, J. E. Labra Gayo, S. Hym, E. G. Prud'hommeaux, H. Solbrig, Complexity and Expressiveness of ShEx for RDF, in: 18th International Conference on Database Theory (ICDT), 2015. doi:10.4230/LIPIcs.ICDT.2015.195.
- [54] K. Thornton, H. Solbrig, G. S. Stupp, L..., A. Waagmeester, Using shape expressions (shex) to share rdf data models and to guide curation with rigorous validation, in: The Semantic Web, Cham, 2019. doi:10.1007/978-3-030-21348-0_39.

- [55] G. Giussani, S. Steinbuss, Data Connector Report, Technical Report, International Data Spaces Association, 2022. URL: https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/Data-Connector-Report-1.pdf, accessed: 2024-02-05.
- [56] M. Atzori, A. Ciaramella, C. Diamantini, B. Martino, S. Distefano, T. Facchinetti, F. Montecchiani, A. Nocera, G. Ruffo, R. Trasarti, et al., Dataspaces: Concepts, architectures and initiatives, in: The 2nd Italian Conference on Big Data and Data Science, volume 3606, CEUR-WS, 2024. URL: <https://hdl.handle.net/11584/389724>.
- [57] O. Corcho, M. Eriksson, K. Kurowski, M. Ojsteršek, C. Choirat, M. Sanden, F. Coppens, EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021. doi:10.2777/620649.
- [58] C. Doukeridis, G. M. Santipantakis, N. Koutroumanis, ..., M. Falsetta, Mobispaces: An architecture for energy-efficient data spaces for mobility data, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023. doi:10.1109/BigData59044.2023.10386539.
- [59] S. Meckler, R. Dorsch, D. Henselmann, A. Harth, The web and linked data as a solid foundation for dataspaces, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 1440–1446. doi:10.1145/3543873.3587616.