Exploring Content-Based Catalogs for Enhanced Discovery Services in Data Spaces

Adriana Morejón[1,*], Alberto Berenguer[1], Lucía de Espona[1], David Tomás[1] and Jose-Norberto Mazón[1]

[1] University Institute for Computing Research, Department of Software and Computing Systems, University of Alicante, Spain

Abstract

In the realm of data sharing, effective data discovery is critical for fostering collaboration and innovation across organizations. Data spaces merge the interoperability and secure sharing features of data ecosystems with the transactional and economic aspects of data markets, enabling seamless data exchange among data providers and consumers while preserving data sovereignty. Central to this paradigm is the data catalog, an entity that handles metadata and provides it for data discovery services. However, traditional data catalogs rely heavily on metadata related to high-level representation of data, as well as the overall features of datasets (e.g., dataset name, license or keywords, following standards such as DCAT). This limitation hampers data discovery, as these metadata alone may not effectively describe datasets to fully convey the relevance of datasets. To address this challenge, this paper proposes content-based catalogs to enhance data discovery within data spaces. Our approach for content-based catalogs incorporates three key assets for data consumers to discover relevant datasets and evaluate their utility before accessing them: (i) high-quality structural metadata for datasets; (ii) representative data samples from the dataset; and (iii) a discovery service for searching relevant datasets. Our novel content-based catalog reduces the risk of unintended exposure while still showcasing the value of data, thus preserving the interests of both data providers and consumers within data spaces.

Keywords

data space, data discovery, data catalog, data samples, metadata

1. Introduction

In today's data-driven landscape, the effective management and exchange of data are central to both organization strategy and innovation. Data ecosystems and data markets, in particular, represent critical frameworks for facilitating data exchange between participants. While data ecosystems focus on allowing organizations to share data in secure environments with an emphasis on interoperability and privacy, data markets are transactional platforms where data is monetized. Data spaces emerge from the confluence of data ecosystems and data markets [1]. They leverage the collaborative and interoperable nature of ecosystems to ensure seamless data sharing among participating organizations, while also enabling structured transactions and value exchanges, as in data markets. This combined approach responds to both the strategic and operational data needs and offers of participants (data consumers and data providers), fostering a dynamic scenario where data must be discoverable before being reused in a variety of applications [2], while data sovereignty is preserved [3]. This brings attention to the critical role of data discovery in data spaces, as data consumers must identify the appropriate datasets from data providers to develop products and services. This necessity has driven a growing focus on creating innovative mechanisms that enable the efficient and effective discovery of relevant data [4]. Traditionally, data discovery and exploration rely on metadata provided from data catalogs to help data consumers find relevant data for their needs [5], even within enterprises [6, 7].

Data catalogs in data spaces are typically based on DCAT standard [8], which have proven effective in the context of open data. DCAT metadata includes high-level descriptors of datasets, such as title, description, keywords, identifier, classification, publisher, license, and technical metadata (e.g., access URLs, format, and size). Although DCAT metadata facilitates the retrieval of basic data sets on open data portals, they are inherently limited to address more complex data discovery needs [9]. This limitation can hinder the identification of datasets that precisely meet specific requirements or use cases, particularly in complex or domain-specific scenarios such as data spaces [10]. Data catalogs for data spaces must semantically describe data content and structure of data, thus allowing data consumers to efficiently discover, understand, and assess the relevance of available data for specific use cases [11]. In order to preserve organizational data sovereignty, current approaches for data space catalogs offer metadata as semantic descriptions of available data sources without offering the content itself [12].

Interestingly, leveraging the content of datasets themselves can enhance discovery processes by enabling data consumers to evaluate data and verify alignment with their needs. However, these content-based catalogs pose a significant challenge to data spaces due to the Arrow paradox [13]. This economic theory highlights the inherent difficulty of valuing data before it is shared because, once the data is disclosed, its value is essentially transferred to the recipient. If content-based catalogs are used in data spaces for improving discoverability, this creates a dilemma for data providers: to attract potential consumers, they must reveal enough about their datasets to demonstrate value, but doing so risks exposing data in a way that diminishes its sovereignty. Similarly, data consumers face uncertainty about the usefulness of data before obtaining it, making it challenging to justify investment or commitment in the data space.

Therefore, novel content-based catalog solutions must balance the need for richer discovery mechanisms with the constraints of data sovereignty. This requires a new generation of data catalogs capable of managing two levels of content-based metadata: (i) descriptions of dataset content derived from its internal structure (e.g., field names and descriptions), and (ii) controlled, privacy-preserving data samples that facilitate dataset evaluation without disclosing all the data.

To face this challenge, this paper proposes content-based

DOLAP 2025: 27th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2025, March 25, 2025, Barcelona, Spain

*Corresponding author.

0009-0005-1124-9682 (A. Morejón); 0000-0002-2867-8329 (A. Berenguer); 0000-0002-1477-6999 (L. d. Espona); 0000-0003-3287-9366 (D. Tomás); 0000-0001-7924-0880 (J. Mazón)

catalogs that include: (i) structural metadata by using Datapackage of Frictionless Data project [14]; (ii) data samples as the subset of data that behaves in the same way that the original dataset for data discovery [15], as well as (iii) a discovery service [16] built upon a metadata repository harvesting the above information across the multiple data providers of the data space. Our novel catalog for data sapces streamlines the data discovery process for consumers, assessing the relevance of data without fully exposing its content until agreements for sharing data between consumers and providers are in place.

The remainder of this paper is structured as follows: Section 2 presents related work to the field of data spaces architectures; Section 3 describes our architecture for data spaces that considers a novel content-based catalog for data discovery services; content-based catalog for data spaces is described in Section 4; finally, Section 5 sketches out conclusions and future work.

2. Related Work

A data space is a federated data infrastructure that supports trustworthy data sharing among data providers and data consumers [17]. To enable this, data spaces rely on connectors [18], which act as interfaces between participants and the ecosystem. Connectors ensure technical interoperability and enforce access policies, allowing providers to keep data sovereignty when sharing data while granting consumers seamless access without compromising control. Hence, data spaces enable new options for value creation where providers and consumers easily interact and work together to find, access, publish, consume, and reuse data, as well as to stimulate innovation [19].

Numerous research initiatives have resulted in conceptual frameworks and reference architecture models to accelerate the development of data spaces [20].

The i4Trust initiative emerges as a collaboration program, targeting the creation of data spaces by proposing an architecture based on commonly agreed building blocks [21]. They are a combination of components from FIWARE and iSHARE, two data space foundations. The iSHARE Foundation [22] maintains a trust framework for data spaces. With the iSHARE Trust Framework, base components are available for data spaces, aligned with the European Strategy for Data[1] and reference architectures like the International Data Spaces Association (IDSA) and Gaia-X described below. The convergence of these architectures enable a unified data platform, contributing to a digital maturity model for building a data ecosystem that strives for standardization. The FIWARE Foundation [23] introduces an open source framework that promotes the development of interoperable, smart solutions.

Another innovative data space technical framework has been developed by Gaia-X, aiming to establish a trustworthy ecosystem where data is shared, maintaining the user's digital sovereignty of the data. This standard-based framework allows the implementation of distributed data systems in all European countries in a legally secure manner, enabling compliance with GDPR and other data regulations [24].

The International Data Spaces Association (IDSA) is a non-profit organization focusing on establishing and promoting standards for data spaces as trusted environments where organizations can share data while retaining full control over its use [25]. The IDS Reference Architecture Model (IDS-RAM) [26], materializes the standard developed from collecting requirements from various industries and the results gained from the model's implementation. Additionally, the IDS-RAM connectors fit the GAIA-X principles and architecture model [27]. IDSA-RAM is suitable for industrial applications, as shown by the architecture developed by the Fraunhofer-Gesellschaft [28, 1, 17].

One of the core building blocks of those reference architectures is the data catalog, which serves as a structured inventory of available datasets and associated metadata and access policies. These catalogs enable data discovery while preserving the sovereignty and privacy of the underlying datasets. Unfortunately, current reference architectures incorporate catalogs that are insufficient for effective data discovery, posing two key limitations that hinder data consumers from accurately evaluating relevance of shared datasets:

- Limited metadata: existing data catalogs primarily rely on high-level metadata to describe datasets but lack detailed structural information of dataset content, such as field names or descriptions, which are essential for precise dataset assessment.
- Content blindness: since federated environments must preserve data sovereignty, current data catalogs exclude actual dataset content, relying solely on metadata. This prevents data consumers from gaining deeper insights into dataset relevance before access is granted.

To overcome these challenges in data discovery, reference architectures for data spaces require complementing data catalogs with content-based metadata while protecting data sovereignty. In response to these requirements, the following section proposes an evolution of a reference architecture that integrates content-based metadata and data discovery mechanisms that respect data sovereignty.

3. Architecture

From the reference architectures described in the previous section, we have chosen the one proposed by IDSA as it has shown compatibility with others. Additionally, it provides an open source ready-to-use Docker-based implementation of a Minimally Viable Data Space (IDS-MVDS) [29]. The IDS-MVDS consists of the following elements:

- **Identity Provider**: the IDS offers identity management across participants of the data space, according to modern standards with low organizational hurdles. Two elements compose the Identity provider in the IDS-MVDS:
  - **Certificate Authority (CA)**: responsible for issue and manage technical identity claims.
  - **Dynamic Attribute Provisioning Service (DAPS)**: provides short-lived tokens with up-to-date information about connectors.
- **Metadata Broker**: The Metadata Broker contains an endpoint for the registration, publication, maintenance, and query of Self-Descriptions from the data providers.
- **Data consumers and data providers**: IDS connectors that request or offer data within the data space. A connector can be simultaneously providing its own data and consuming from another connector.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

The IDS Connectors metadata and data in the IDS ecosystem are structured hierarchically according to the data model based on the structure of the IDS information model. The main entity of this data model is the *resource* as it contains the core metadata of a data object, including the title, description and license information. Related groups of resources are organized by *catalogs*, the top entity in the data model. A resource also has a list of *representations* that describes the format of the offered data. Below the representation, there is the *artifact* that has a 1:1 relation to the raw data and includes low level information such as checksum and byte size. Each artifact has a reference to *contract agreements*, which describe the agreed usage between data provider and data consumer. Contract offers can contain multiple *rules* representing IDS Usage Control Patterns. Further details about the data model entities and their attributes are available at [30].

Among the previously listed IDS-MVDS components, the Metadata Broker plays a crucial role in aggregating metadata that describes the variety of connectors and catalogs in the entire data space ecosystem. Serving as a central repository, the broker primarily focuses on receiving metadata from data providers and delivering it to data consumers within the data space. The broker offers search functionalities but only contains metadata up to the catalog level which is insufficient for a potential data consumer to discover relevant data in the data space. Furthermore, it will require the consumer to access each suitable data provider catalog to request detailed metadata to make an informed decision about starting the negotiation for a particular data offering.

The effectiveness of the metadata broker's search service largely depends on the accuracy and completeness of the metadata provided by data providers at the resource level. However, because these metadata entries are often manually curated, they may lack sufficient detail to fully describe the dataset, limiting the precision of search and retrieval processes. Additionally, data space catalogs do not provide metadata at the artifact level or dataset samples, preventing users from evaluating dataset relevance before access is granted. Unlike open data portals, where datasets can be freely downloaded for assessment, data spaces impose stricter access controls, making discovery more challenging.

Our proposed solution overcomes the limitations of the IDS-MVS architecture by introducing a discovery service that leverages an enhanced data provider catalog with content-based metadata at artifact level, i.e. structural metadata describing data schema (field names, data types, and descriptions of the fields), as well as sample data instances. These content-based metadata is extracted from the datasets and combines it with the eventually existing metadata at resource level (title, keywords, etc.) to generate a comprehensive description of the data offering. This information is then integrated into the data provider catalog to make it available for the data consumers. Specifically, for each offered dataset, an additional sample resource is generated automatically and added to the catalog. This sample resource contains both structural metadata, and a representative subset of the full dataset. The original resource contains a link to this sample resource, following the IDS metadata standard which contains a specific element with this aim. This sample resource is stored locally on the data provider for a simplified retrieval as shown in Figure 1.

These enhancements to the data space reference architecture empower potential data consumers with detailed insights into a dataset's structure and representative samples before initiating full access negotiations. This is achieved through advanced data discovery services. In this context, we introduce InferIA [16], an search engine, based on artificial intelligence, that utilizes our content-based catalog to enhance search accuracy and relevance, outperforming traditional metadata brokers. To connect this service with the IDS data space ecosystem, we have developed the following suite of components depicted on Figure 1:

- **Data consumer**: essential connector for the discovery service that allows the browser to act as a data consumer and collect metadata from the data space.
- **Browser**: the browser component queries periodically the Metadata Broker through the data consumer API to obtain the list of data providers and their catalogs. For each of those catalogs, it requests a list of offered resources and all the available metadata at every entity level including the providers' data catalogs, offered resources and artifacts. The linked sample resources are also requested to obtain their associated content-based metadata at artifact level. The browser dumps the collected data on a temporary database available for the search engine.
- **Metadata storage**: our approach utilizes an ElasticSearch database to structure and optimize the data collected by the browser, as well as to store the word embeddings generated by the embedding microservice (described below). This setup facilitates efficient retrieval of relevant data, making it easier for data consumers to find what they need.
- **Embedding microservice**: embeddings for datasets are generated using a large language model (LLM). For example, in Spanish, we use OpenAI's *text-embedding-3-small* model.
- **Search microservice**: this component provides the algorithm to retrieve the data from the metadata storage, re-rank, and sort the results from a given query according to our previous work [10, 31, 32].
- **API**: used as a middleware to consume the search microservice.

4. Content-based Catalogs

Our approach of content-based catalog for data spaces adds metadata at both the resource and data artifact levels. At the resource level, the metadata follows the IDS standard, composed of multiple fields describing the entity that has been added by the data provider. This resource-level metadata is integrated with the content-based metadata obtained automatically from the data and formatted using the Data Package standard from Frictionless Data project [14] and then combined together with the data sample to be stored locally in the data provider at the artifact level (as shown in Figure 1). This way, the content-based metadata will be obtained together with the data sample when requesting the sample resource.

The Data Package standard is a simple container format for describing a coherent collection of data in a single package. It provides the basis for the convenient delivery, installation, and management of datasets [33]. The metadata within the Data Package that applies to resource level is the same as the one included in the sample resource at a higher level of IDS entity hierarchy adapted to the Data Package

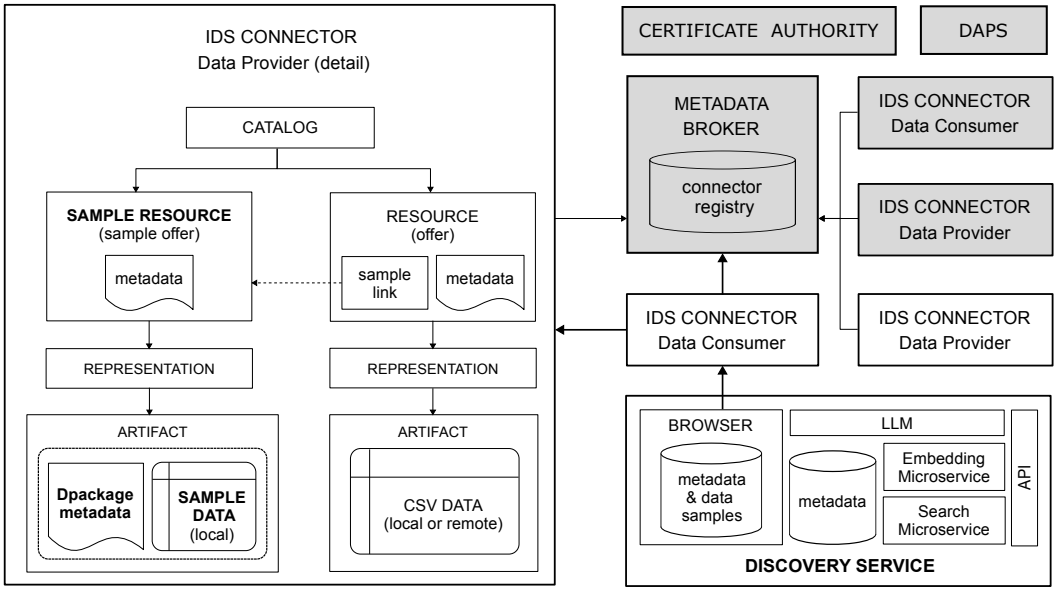

Figure 1: Overview of our extension of the minimally viable data space proposed in the IDS Testbed. Two extra data providers were added as well as a data consumer to collect metadata for the discovery service by browsing through the data catalogs. The part marked in gray corresponds to the IDS-MVDS.

standard. Both of them are compatible with DCAT W3C standard [8], ensuring interoperability.

Furthermore, additional metadata on schema of the data is included: the name of each field, the description of each field, the data type, and a sample value. This information can be added for each field by using Frictionless [14].

Finally, to efficiently generate representative data samples while maintaining computational efficiency, we employ an approach that uses word embeddings to assess data similarity according to our previous work [15]. Traditional methods for generating semantic representations can be resource-intensive, especially for large datasets. To mitigate this, we hypothesize that a substantial portion of a dataset's content can be removed without significantly affecting its semantic representation. To validate this hypothesis, we evaluated different reduction techniques, demonstrating that retaining only a small percentage of the original data preserves its representational quality while reducing computational costs. Three distinct methods were applied: (i) random reduction, which randomly discards a fixed percentage of data; (ii) duplicate removal, which eliminates redundant entries to enhance uniqueness; and (iii) TF-IDF selection, which prioritizes the most informative elements based on their statistical significance. Specifically, among the three proposed reduction strategies, TF-IDF demonstrated superior performance overall compared to random reduction and duplicate removal [15]. These techniques not only reduce data volume but also improve the performance of search and retrieval systems, enabling more efficient and scalable data discovery within the data space.

5. Conclusions and Future Work

In this paper, we propose a novel architecture for data spaces designed to integrate content-based metadata generation and discovery mechanisms, ensuring both compliance with data sovereignty principles and enhanced effectiveness in identifying relevant datasets within the data space ecosystem. We addressed the challenges of improving data discovery in data spaces by proposing an innovative approach that combines descriptive metadata from data schema with data sampling. While traditional metadata catalogs serve as an essential foundation for organizing and describing datasets, their limitations in conveying dataset-specific details often impede effective data discovery. Our approach mitigates these limitations by leveraging both, structural metadata and data sampling techniques that balance data value demonstration with the protection of data sovereignty. By empowering data consumers to assess dataset relevance while safeguarding data provider sovereignty, our approach establishes a significant step toward operationalizing effective data discovery in federated environments, such as data spaces. On the basis on this research for improving data discovery in data spaces, there are several promising directions for future work. Adaptive sampling methodologies that can dynamically adjust to the characteristics of diverse datasets across data providers is an important next step. This could involve leveraging advanced clustering algorithms, domain-specific heuristics, or adaptive machine learning models to optimize the sampling process. Also, ensuring that the sampling techniques align with privacy requirements is crucial. Future work could also explore integrating privacy-preserving mechanisms to enhance trust and compliance with legal and ethical standards. Finally, evaluation of content-based catalogs in various real-world sectoral data spaces, such as healthcare or environment, would provide insights into its adaptability to specific domains.

Acknowledgments

This work is part of the project TED2021-130890B-C21, funded by MCIN/AEI/10.1 3039501100011033 and by the European Union NextGenerationEU/PRTR.

References

[1] F. Möller, I. Jussen, V. Springer, A. Gieß, J. C. Schweihoff, J. Gelhaar, T. Guggenberger, B. Otto, Industrial data ecosystems and data spaces, Electronic Markets 34 (2024) 1–17.

[2] R. C. Fernandez, P. Subramaniam, M. J. Franklin, Data market platforms: trading data assets to solve data problems, Proceedings of the VLDB Endowment 13 (2020) 1933–1947.

[3] P. Hummel, M. Braun, M. Tretter, P. Dabrock, Data sovereignty: A review, Big Data & Society 8 (2021) 2053951720982012.

[4] A. Bhandari, A. Fariha, B. Price, A. Vanterpool, A. Bowne, L. McEvoy, V. Gadepally, et al., Examples are all you need: Iterative data discovery by example in data lakes, in: CIDR, 2022.

[5] C. Gröger, There is no AI without data, Communications of the ACM 64 (2021) 98–108.

[6] R. Eichler, C. Gröger, E. Hoos, H. Schwarz, B. Mitschang, Data shopping—how an enterprise data marketplace supports data democratization in companies, in: International Conference on Advanced Information Systems Engineering, Springer, 2022, pp. 19–26.

[7] C. Labadie, C. Legner, M. Eurich, M. Fadler, Fair enough? enhancing the usage of enterprise data with data catalogs, in: 2020 IEEE 22nd Conference on Business Informatics (CBI), volume 1, IEEE, 2020, pp. 201–210.

[8] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, P. Winstanley, The w3c data catalog vocabulary, version 2: Rationale, design principles, and uptake, Data Intelligence 6 (2024) 457–487.

[9] P. Křemen, M. Nečaskỳ, Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary, Journal of Web Semantics 55 (2019) 1–20.

[10] A. Berenguer, D. Tomás, J. Mazón, Tabular open government data search for data spaces based on word embeddings, in: E. Gallinucci, L. Golab (Eds.), Proceedings of the 25th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) co-located with the 26th International Conference on Extending Database Technology and the 26th International Conference on Database Theory (EDBT/ICDT 2023), Ioannina, Greece, March 28, 2023, volume 3369 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 61–70. URL: https://ceur-ws.org/Vol-3369/paper6.pdf.

[11] M. Hauff, L. M. Comet, P. Moosmann, C. Lange, I. Chrysakis, J. Theissen-Lipp, FAIRness in dataspaces: The role of semantics for data management, in: The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, 2024.

[12] N. Jahnke, B. Otto, Data catalogs in the enterprise: applications and integration, Datenbank-Spektrum 23 (2023) 89–96.

[13] S. A. Azcoitia, N. Laoutaris, A survey of data marketplaces and their business models, ACM SIGMOD Record 51 (2022) 18–29.

[14] D. Fowler, J. Barratt, P. Walsh, Frictionless data: making research data quality visible, International Journal of Digital Curation 12 (2017) 274–285.

[15] A. Berenguer, D. Tomás, J. Mazón, Evaluating the impact of content deletion on tabular data similarity and retrieval using contextual word embeddings, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II, volume 14609 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 433–447. URL: https://doi.org/10.1007/978-3-031-56060-6_28. doi:10.1007/978-3-031-56060-6_28.

[16] A. Berenguer, O. Alcaraz, D. Tomás, J. Mazón, From research on data-intensive software to innovation in data spaces: A search service for tabular data, IEEE Softw. 41 (2024) 59–66. URL: https://doi.org/10.1109/MS.2024.3359333. doi:10.1109/MS.2024.3359333.

[17] B. Otto, A federated infrastructure for european data spaces, Communications of the ACM 65 (2022) 44–45.

[18] A. Gieß, M. J. Hupperz, T. Schoormann, F. Möller, What does it take to connect? unveiling characteristics of data space connectors, in: T. X. Bui (Ed.), 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, January 3-6, 2024, ScholarSpace, 2024, pp. 4238–4247. URL: https://hdl.handle.net/10125/106895.

[19] A. Gieß, F. Möller, T. Schoormann, B. Otto, Design options for data spaces, in: Thirty-first European Conference on Information Systems (ECIS 2023), 2023.

[20] M. Bacco, A. Kocian, S. Chessa, A. Crivello, P. Barsocchi, What are data spaces? systematic survey and future outlook, Data in Brief 57 (2024) 110969.

[21] i4Trust, B2B Data Sharing Playbook - the i4Trust approach to Data Sharing, 2021. URL: https://i4trust.org/wp-content/uploads/i4Trust_DataSharingPlaybook.pdf.

[22] iSHARE Foundation, iSHARE - Trust Framework for Data Spaces, 2021. URL: https://ishare.eu/.

[23] FIWARE Foundation, FIWARE - Open APIs for Open Minds, 2024. URL: https://www.fiware.org.

[24] Gaia-X European Association for Data and Cloud AISBL, Gaia-X Architecture Document, 2024. URL: https://docs.gaia-x.eu/technical-committee/architecture-document/latest/.

[25] International Data Spaces Association, International Data Spaces Association, 2022. URL: https://internationaldataspaces.org/.

[26] International Data Spaces Association, IDS Reference Architecture Model, Version 4.0, 2022. URL: https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/introduction/1_1_goals_of_the_international_data_spaces.

[27] International Data Spaces Association, Position Paper - GAIA-X and IDS, 2021. URL: https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-GAIA-X-and-IDS.pdf.

[28] Fraunhofer-Institut für Software- und Systemtechnik ISST, Fraunhofer - Software(-architecture), 2024. URL: https://www.dataspaces.fraunhofer.de/en/software.html.

[29] International Data Spaces Association, Idsa github repository, 2024. URL: https://github.com/International-Data-Spaces-Association.

[30] International Data Spaces Association, Idsa

data model documentation, 2024. URL: https://international-data-spaces-association.github.io/DataspaceConnector/Documentation/v5/DataModel.

[31] J. Pilaluisa, D. Tomás, B. Navarro-Colorado, J. Mazón, Contextual word embeddings for tabular data search and integration, Neural Comput. Appl. 35 (2023) 9319–9333. URL: https://doi.org/10.1007/s00521-022-08066-8. doi:10.1007/S00521-022-08066-8.

[32] A. Berenguer, J. Mazón, D. Tomás, Word embeddings for retrieving tabular data from research publications, Mach. Learn. 113 (2024) 2227–2248. URL: https://doi.org/10.1007/s10994-023-06472-0. doi:10.1007/S10994-023-06472-0.

[33] Data Package Working Group, Data Package Standard, 2024. URL: https://datapackage.org/standard/data-package/.