Dataspaces: concepts, architectures and initiatives

Maurizio Atzori[a,1], Angelo Ciaramella[a,2], Claudia Diamantini[a,3], Beniamino Di Martino[a,4], Salvatore Distefano[a,5], Tullio Facchinetti[a,6], Fabrizio Montecchiani[a,7], Antonino Nocera[a,6], Giancarlo Ruffo[a,8], Roberto Trasarti[a,9]

[a] *Consorzio CINI - Laboratorio Data Science*
[1] *Università degli Studi di Cagliari, Via Ospedale 72, 09124 Cagliari, Italia - atzori@unica.it*
[2] *Università degli Studi di Napoli "Parthenope", DiST, Centro Direzionale di Napoli, Isola C4, 80143, Napoli, Italia - angelo.ciaramella@uniparthenope.it*
[3] *Università Politecnica delle Marche, via Brecce Bianche, Ancona, Italia - c.diamantini@univpm.it*
[4] *Università degli Studi della Campania "Luigi Vanvitelli" - beniamino.dimartino@unicampania.it*
[5] *Università degli Studi di Messina, Piazza Pugliatti 1, 98100 Messina, Italia - sdistefano@unime.it*
[6] *Università degli Studi di Pavia, DIII, Via A. Ferrata 5, 27100 Pavia, Italia - antonino.nocera@unipv.it; tullio.facchinetti@unipv.it*
[7] *Università degli Studi di Perugia, DI, Via G. Duranti 93, 06125, Perugia, Italia - fabrizio.montecchiani@unipg.it*
[8] *Università degli Studi del Piemonte Orientale "A. Avogadro", Via Duomo, 6 - 13100 Vercelli, Italia - giancarlo.ruffo@uniupo.it*
[9] *Consiglio Nazionale delle Ricerche - ISTI - KDD Lab, 56124 Via G.Moruzzi 1, Pisa, Italia - roberto.trasarti@isti.cnr.it*

Abstract

Despite not being a new concept, dataspaces have become a prominent topic due to the increasing availability of data and the need for efficient management and utilization of diverse data sources. In simple terms, a dataspace refers to an environment where data from various sources, formats, and domains can be integrated, shared, and analyzed. It aims to provide a unified view of heterogeneous data by bridging the gap between different data silos, enabling interoperability. The concept of dataspaces promotes the idea that data should be treated as a cohesive entity, rather than being fragmented across different systems and applications.

Dataspaces often involve the integration of structured and unstructured data, including databases, documents, sensor data, social media feeds, and more. The goal is to enable organizations to harness the full potential of their data assets by facilitating data discovery, access, and analysis. By bringing together diverse data sources, dataspaces can offer new insights, support decision-making processes, and drive innovation.

In the context of European Commission-funded research projects, dataspaces are often explored as part of initiatives focused on data management, data sharing, and the development of data-driven technologies. These projects aim to address challenges related to data integration, data privacy, data governance, and scalability. The goal is to advance the state of the art in data management and enable organizations to leverage data more effectively for societal, economic, and scientific advancements.

It is important to notice that while dataspaces offer potential benefits, they also come with challenges. These challenges include data quality assurance, data privacy and security, semantic interoperability, scalability, and the need for appropriate data governance frameworks.

Overall, dataspaces represent an approach to managing and utilizing data that emphasizes integration, interoperability, and accessibility. The concept is being explored and researched to develop innovative solutions that can unlock the value of data in various domains and sectors.

Keywords: Dataspaces, Data Lakes, Data Integration

1. Introduction and Main Concepts

The term dataspace (DS) was originally coined by Franklyn and other authors [Franklin et. al. 2005, Halevy et al. 2006] as an evolution of traditional DBMS. Since the introduction of the definition, several other references to this concept have been elaborated in the scientific literature. The rather vague nature of the original definition led to slightly different semantics and variations in the references appeared later.

In the original concept, a dataspace is defined as an "abstraction" for data management focused on reducing the challenges behind the fruitful and efficient exploitation of large amounts of interrelated, although disparately managed, data. Along with this definition, the authors of the aforementioned manuscript came up with a second, and perhaps more important, concept, namely the Dataspace Support Platform (DSSP, for short).

According to their vision, in a context in which an increasing number of loosely linked data sources can be leveraged to feed novel and advanced application scenarios, the challenges related to data management become pervasive as they have to be solved on every single source, individually. For this reason, the classical concepts of databases and Database Management Systems (DBMS, for short) should be replaced by more abstract definitions.

In this sense, a dataspace can be thought of as just an abstract database, whose data are actually located in independent and heterogeneous data platforms, possibly sharing common semantics. Semantic integration, typically required in classical DBMS, is also being released and, according to [Franklin et. al. 2005, Halevy et al. 2006], dataspaces and DSSP should provide more of a "data coexistence" strategy than full data integration. Roughly speaking, dataspaces and DSSP should provide basic data access capabilities across different data sources and, therefore, implicitly defer the definition of fine-tuned integration policies to the application layer.

This initial definition has been extended over the years in several directions. The concept of data space has seen renewed practical interest due to the EU initiative of European Common Dataspaces[1], aiming to enforce data sovereignty and establish a data economy. Recently, [Curry, 2020] identified different features to complete the definition of a dataspace, which have been extended here to the 13 ones reported below:

1. Storage Architecture: Refers to the underlying structure and organization of data storage, distinguishing between centralized (data stored in a single location) and distributed (data stored across multiple locations) architectures.
2. Control: Describes the level of centralization or distribution of control over data management, ranging from centralized-complete (single entity controls all aspects) to distributed-partial (multiple entities have autonomy over specific aspects).
3. Model: Represents the data model or database model used to structure and organize data, such as relational, NoSQL, or hybrid models.
4. Formats: Refers to the types of data formats supported, including structured (data organized in a predefined format), semi-structured (data with a loose structure like JSON or XML), unstructured (data without a predefined structure, such as text documents or multimedia).
5. Schema: Defines how data schema (structure and organization) is handled, including schema-first (schema defined before data is stored), data-first (data stored without a predefined schema, later defined), or no schema (data stored without any predefined structure).
6. Integration: Describes the approach to integrating data from different sources, ranging from upfront-strong integration (data integration before storage), incremental-weak integration (periodic or incremental integration), to on-demand-none integration (no predefined integration process).

[1] Communication: A European Strategy for Data, 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

7. Leadership: Refers to the leadership approach in managing the data management system, distinguishing between top-down (centralized decision-making) and bottom-up (distributed decision-making) approaches.
8. Query: Defines the type of queries supported by the system, including exact queries (precise retrieval of specific data) and approximate queries (approximate or probabilistic retrieval).
9. Data Processing: Refers to the methods and capabilities for processing data, such as real-time streaming (processing data as it arrives), batch processing (processing data in large batches), or other data processing approaches.
10. Governance: Describes the data governance model applied to data management, including centralized governance (centralized control and decision-making), distributed governance (distributed control and decision-making), or a combination of both.
11. Sovereignty: Represents the level of data ownership and control, ranging from none (no individual control over data), weak (partial control) to full-strong sovereignty (complete individual control over data).
12. Trustworthiness: Indicates the level of trust and reliability in the data management system, ranging from none (lack of trustworthiness), weak, to strong trustworthiness (highly reliable and trusted system).
13. Consistency and Durability: Describes the level of data consistency (ensuring data correctness and integrity) and durability (ensuring data persistence and availability) provided by the system, ranging from none (lack of consistency and durability), weak, to strong (highly consistent and durable system).

1.1 Current Challenges

In the context of European Data Spaces, a number of issues have been identified by BDVA that need to be addressed to make dataspaces effective [Scerri et al., 2022].
They can be clustered into 4 groups of challenges that consider different aspects of running a dataspace: technical challenges, business and organizational challenges, legal compliance challenges, national and regional challenges. Focusing on the technical challenges, the main problems that need to be addressed are:

- *Sharing by Design*: a dataspace should have a data lifecycle management model that includes sharing by design, that is conceived to facilitate the sharing of interoperable data and provide mechanisms to integrate them
- *Digital Sovereignty*: new ownership models or appropriate tools for data rights management need to be developed to enforce data usage rights within a mixed data sharing space such as those made available by dataspaces.
- *Decentralization*: it is challenging to guarantee scalability of real-time data operations in massively distributed data architectures whose distribution is not defined apriori.
- *Veracity*: dataspaces need tools for verification and provenance support given their data sharing nature
- *Security*: secure data access and restrictions policies should take into account the sharing goal of dataspaces, which make it challenging to ensure confidentiality and digital rights management compared to other data management solutions. Also communication among the nodes of the decentralized architecture would need a secure network and appropriate protocols.
- *Privacy protection*: although privacy-preserving technologies in the context of databases are available, they need to be adapted to address the challenges posed by the way data are shared via dataspaces [Dutkiewicz et al., 2022].

Other important problems that remain open for further research are query performance, which may suffer from missing centralized data indexes or optimized partitioning of data, as well as the application of AI techniques and algorithms in order to automatically construct a mediated schema from various sources [Nargesian et al., 2019; Jarke, et al., 2022], in order to reduce the cost of data integration of the pay as you go paradigm.

1.2 Existing Data Management solutions

To better understand DS, it could be useful to compare such technology against other relevant data management solutions such as data lakes, data warehouses, and different flavors of DBs:

- *Data Lakes*: Data Lakes are storage repositories that store large volumes of raw and unprocessed data in its native format. They provide a centralized location for storing diverse data sources, making it easier to analyze and derive insights. While data lakes focus on storage and provide limited data organization and integration capabilities, data spaces go beyond storage and provide an integrated environment for organizing, exploring, and analyzing data from diverse sources. Data spaces offer user-centric operations and support navigation, search, and exploration through different interfaces.
- *Data Warehouse*: A data warehouse is a centralized repository that consolidates data from various sources for reporting, analysis, and decision-making purposes. It typically involves data integration, transformation, and aggregation processes. Data spaces share some similarities with data warehouses in terms of integrating and organizing data from multiple sources. However, data spaces focus on providing a user-centric environment for exploring and analyzing data, while data warehouses primarily focus on supporting business intelligence and reporting.
- *Databases (DB)*: Databases are structured collections of data organized for efficient storage, retrieval, and management. Data spaces can incorporate databases as one of the data sources within their environment. However, data spaces typically go beyond individual databases and provide a unified view that integrates data from multiple sources, including databases, into a cohesive environment.
- *Distributed Database (DDB)*: A distributed database is a database that is spread across multiple nodes or locations, providing improved scalability and fault tolerance. Data spaces can integrate data from distributed databases as part of their data sources. However, data spaces go beyond distributed databases by providing a unified and virtualized environment that integrates data from various sources, regardless of their distribution or location.
- *Federated Database (FDB)*: A federated database is a collection of autonomous databases that are interconnected and present a unified view to users. Federated databases allow querying and accessing data from multiple databases through a single interface. Data spaces, similar to federated databases, integrate data from multiple sources. However, data spaces offer additional capabilities such as exploration, navigation, and user-centric operations that enhance the user's experience with the integrated data.
- *Multi-Database (Multi-DB)*: A multi-database system consists of multiple independent databases that operate concurrently but are not necessarily interconnected. Each database maintains its own data and schema. In contrast, data spaces provide an integrated environment that bridges the gap between multiple databases, allowing users to work with data from different sources seamlessly.

Restricting the scope to the two first dimensions (storage architecture and control), a taxonomy able to give a preliminary categorization and positioning of the above discussed systems in the data management technology landscape is proposed below.

Architecture/CTRL	Centralized	Distributed
Centralized	Data Lake	Data Warehouse
Distributed	Distributed DB, Distributed/Cloud FS	Dataspace/ Federated-MultiDB

Table 1 summarizes the full comparison among the aforementioned data management solutions based on all the features above identified. In particular, there is an even distribution among centralized and distributed solutions. Moreover, database solutions are based on SQL/NoSQL approaches, leading to structured schemes, which are not required or addressed in dataspaces, data lakes or data warehouses. The level of trustworthiness, consistency and durability spans from weak to strong, where several solutions can be configured to obtain the desired level of each feature. Dataspaces set apart from all the other solutions regarding the type of query, which is exact for all the solutions but can be approximated for dataspaces, and for the leadership, which is top-down for all the solutions.

Feature	**Data Spaces**	**Data Lakes**	**Data Warehouses**	**Databases (DB)**	**Distributed Databases (DDB)**	**Federated Databases (FDB)**	**Multi-Databases (Multi-DB)**
Storage Architecture	Distributed	Centralized	Centralized	Centralized	Distributed	Distributed	Distributed
Control	Distributed	Centralized	Centralized	Centralized	Centralized	Distributed	Distributed
Model	*	*	*	SQL/ NoSQL	SQL/NoSQL	SQL/NoSQL	SQL/NoSQL
Formats	*	*	*	Structured	Structured	Structured	Structured
Schema	Data-first/ no schema	*	*	Schema first	Schema first	Schema first	Schema first
Integration	Weak or incremental	*	Upfront-strong	Upfront-strong	Upfront-strong	Upfront-strong	Upfront-strong
Leadership	*	Top-down	Top-down	Top-down	Top-down	Top-down	Top-down
Query	Exact, Approx.	Exact	Exact	Exact	Exact	Exact	Exact
Data Processing	*	Batch	Batch, Near-real-time	Batch	*	*	*
Governance	*	Centralized	Centralized	Centralized	Centralized	Centralized	Centralized
Sovereignty	*	Partial	Partial/Full	Full	Partial/Full	Partial/Full	Partial/Full
Trustworthiness	Strong	Weak-Strong	Strong	*	*	Weak-Strong	*
Consistency and Durability	Weak	Weak-Strong	Strong	Strong	Weak-Strong	Weak-Strong	Weak-Strong

Table 1. Comparison of the relevant features among different data storage management solutions.

2. Architectural frameworks

Concerning dataspaces, in the European scenario, three main initiatives focus on addressing the challenge of data publishing and sharing. Those initiatives are: EOSC, IDS, DSBA and Gaia-X, all of them are incorporating FAIR principles (findable, accessible, interoperable, and reusable) in both scientific and commercial areas. All those initiatives define their architecture around some basic concepts (such as [IDS-RAM 2019]):

- trust
- security and data sovereignty
- ecosystem of data (no central data storage capabilities)
- standardized interoperability
- value adding apps
- data markets
- open development process

- re-use of existing technologies
- contribution to standardization

EGI foundation produced a paper comparing the approaches of EOSC and Gaia-X [EGI, 2021], in the following we will report a short comparison between the four initiatives.

EOSC is based on a four layer architecture (i) **EOSC-Core** defined by the internal services allowing EOSC to operate as a federation. It includes a Core technical platform which facilitates EOSC delivery upon which the researcher facing resources in the EOSC-Exchange can rely and integrate with as appropriate. (ii) **EOSC-Exchange** provides services and other resources registered into the EOSC to serve the needs of research communities. Generic services and resources which target multiple scientific domains and research communities are identified as Horizontal Services. Resources which target users from a specific scientific domain, community and/or regional domain are identified as Thematic and/or Regional Resources. The capability to compose resources across horizontal and thematic and/or regional ones relies on the EOSC Interoperability Framework. (iii) **EOSC Interoperability Framework (EIF)** is a framework of standards and guidelines to support the interoperability and composability of resources in the EOSC-Core and EOSCExchange. It allows EOSC to integrate services and research products (e.g. publications, datasets, software) across resources and providers. Providers have the freedom to develop and operate provider specific implementations while conforming to the EIF guidelines and standards. Data ecosystems delivering thematic capabilities are independently operated outside EOSC for their reference target groups. (iv) **EOSC Support activities,** alongside the EOSC-Core and EOSC-Exchange, comprise the training, engagement, and other human-centric activities which make EOSC more attractive and easier to use, and help users benefit from it more easily once engaged. They include Training, support and the EOSC Digital Innovation Hub for engagement with the commercial sector.
It is important to notice also the timeline and the relations between initiatives: the European dataspaces were introduced as a parallel initiative to the EOSC by the European community, in practice the output of the EOSC initiative will be used as a base for a fully connected common platform including the dataspaces.

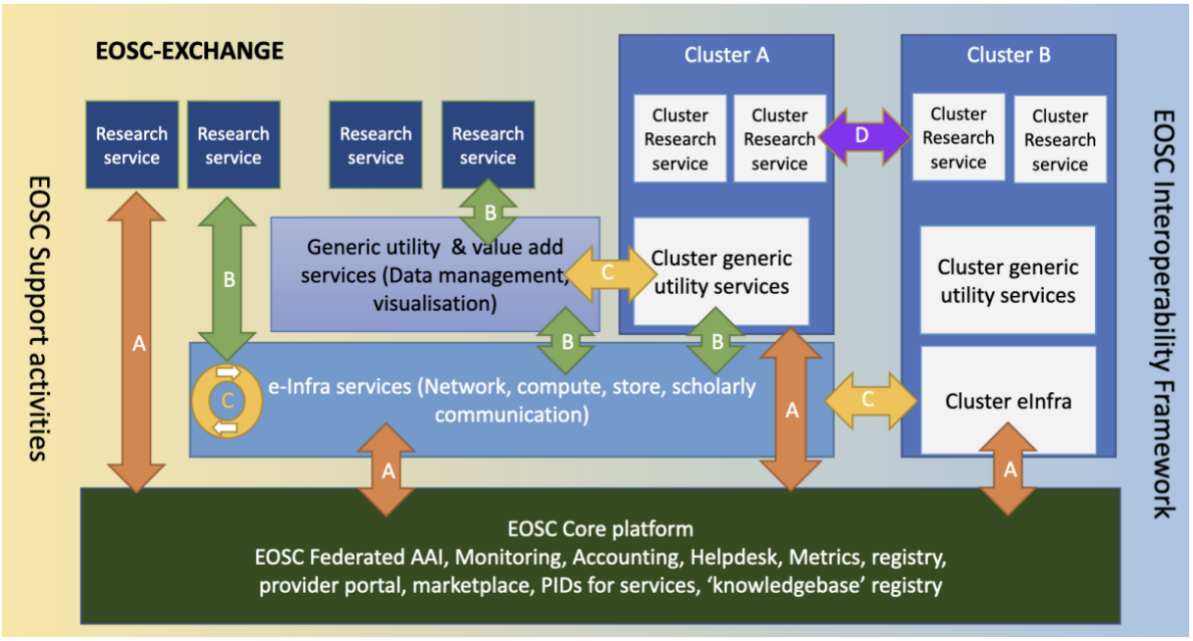

Figure 1: High level architecture of EOSC-Exchange [Licia et al., 2021]

The **Gaia-X** architecture is based on the concepts of Asset and the roles of Data Providers, Federators and Consumers. The design of the system is defined as relations between these concepts: (i) **Asset**: the resources which are shared among the network including meta-data and other information needed for

their usage. (ii) a **Provider** is who provides Assets in the Gaia-X Ecosystem. It defines the service offering including terms and conditions as well as technical policies. further, it provides the service instance that includes a Self-Description and technical policies. Therefore, the Provider may possess different Assets. (iii) **Federators** are in charge of organizing and managing vertical contexts (e.g. similar concept to dataspace) and are autonomous in defining specific rules and policies for asset sharing. (iv) A **Consumer** is a participant who searches and consumes the assets in the Gaia-X ecosystem. The definition of gaia-X architecture is still not well defined in terms of technologies and is more focussed in the creation of a network of trust between industrial partners.

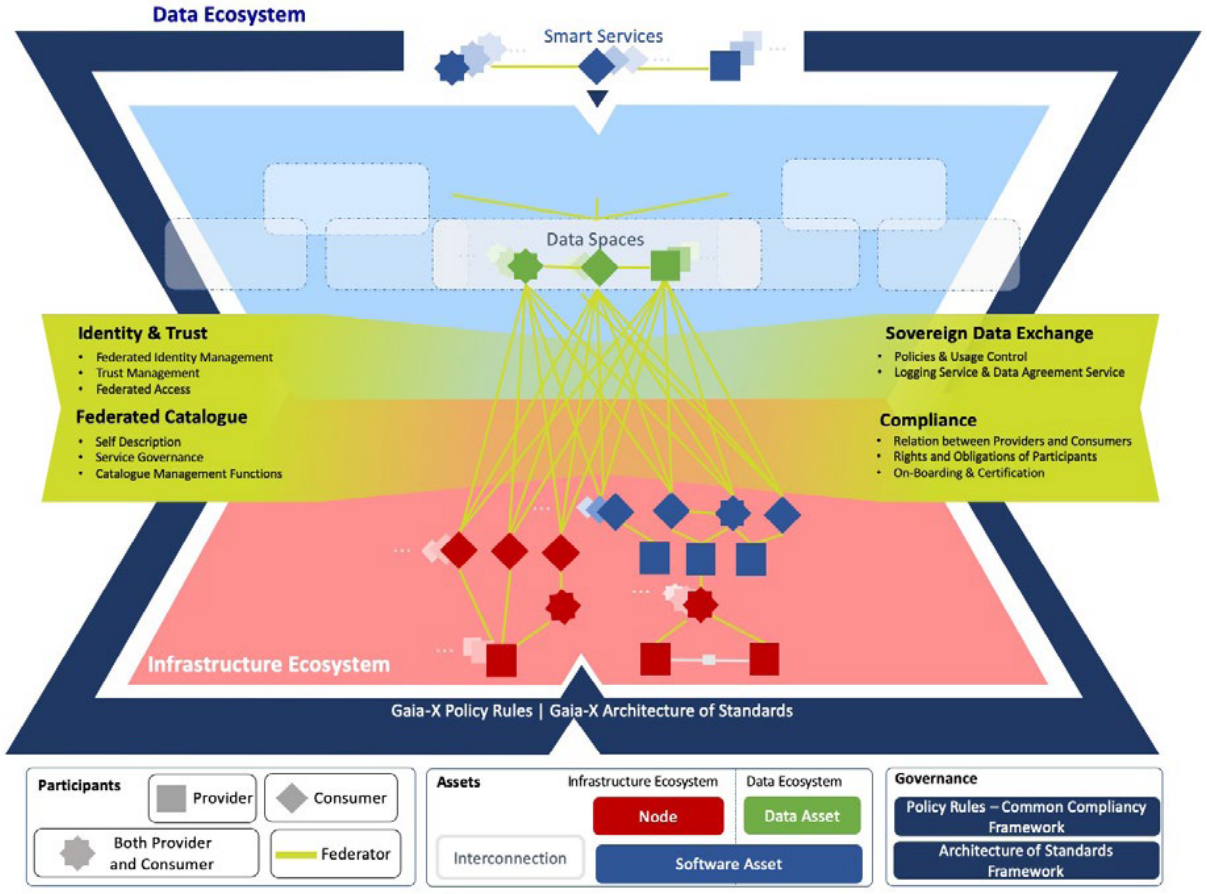

Figure 2: High level architecture of GaiaX initiative[2]

The International Data Space Association (IDS) Reference Architecture Model (IDS-RAM) [IDS-RAM, 2019] is made up of 5 layers and 3 perspectives. The 5 layers describe the structure of Business, Functional, Information, Process, and System components. The 3 perspectives deal with functionalities that have to be implemented across the layers: security, certification and governance.
(i) **The Business layer** provides an abstract description of roles in the International Data Space. and their interactions. It can be considered a blueprint for the other, more technical layers. Principal roles are categorized in Core Participants (e.g. data owner, data provider, data consumer), Intermediary (e.g. broker service provider, clearing house, identity provider) Software/Service Provider, Governance Body (e.g. certification body, IDSA). (ii) **The Functional layer** describes the functional requirements of the IDS, like functionalities to ensure trust (e.g. identity management), security and data sovereignty (e.g. authentication and authorization), Ecosystems of Data (e.g. Data Source description, Brokering, Vocabularies), Standardized Interoperability (e.g. operation, data exchange), Value Adding Apps (e.g. data processing and transformation, data app implementation) Data Markets (clearing and billing, usage restrictions and governance, legal aspects). (iii) **The Process layer** describes, in BPMN notation, the interactions between the different components of the IDS. Three

[2] https://towardsdatascience.com/the-architecture-of-europes-gaia-x-850ba6f43519

major processes have been identified, together with their subprocesses: Onboarding, Exchanging Data, and Publishing and using Data Apps. (iv) **The Information Layer** describes the Information Model. It is defined as an RDFS/OWL-ontology covering the types of Digital Resources that are exchanged by participants by means of the IDS infrastructure components. It supports the description, publication and identification of Digital Resources (both data and data processing software) as well as data exchange and consumption via semantically annotated, easily discoverable services. The framework explicitly assumes that specific domain ontologies and vocabularies can be integrated for more detailed resource annotation. Besides the normative ontology (called the Declarative Representation) the model is specified at two further levels. The former, more abstract, called the Conceptual Representation is a textual document devoted to the general public, complemented with graphical models (UML classes). The latter, more specific, called the Programmatic Representation, provides a mapping of the IDS Ontology onto native structures of a given programming language, targeting Software Providers needs. Finally, the (v) **the System Layer** is the more technical layer, being devoted to map roles (specified on the Business Layer) and requirements (specified on the Functional Layer) onto a concrete architecture. Three major technical components are identified: the Connector, the Broker, and the App Store. Other "external"components (i.e. not specified by IDS-RAM) support the three components: the Identity Provider, The Vocabulary Hub, the Update Repository, and the Trust Repository.

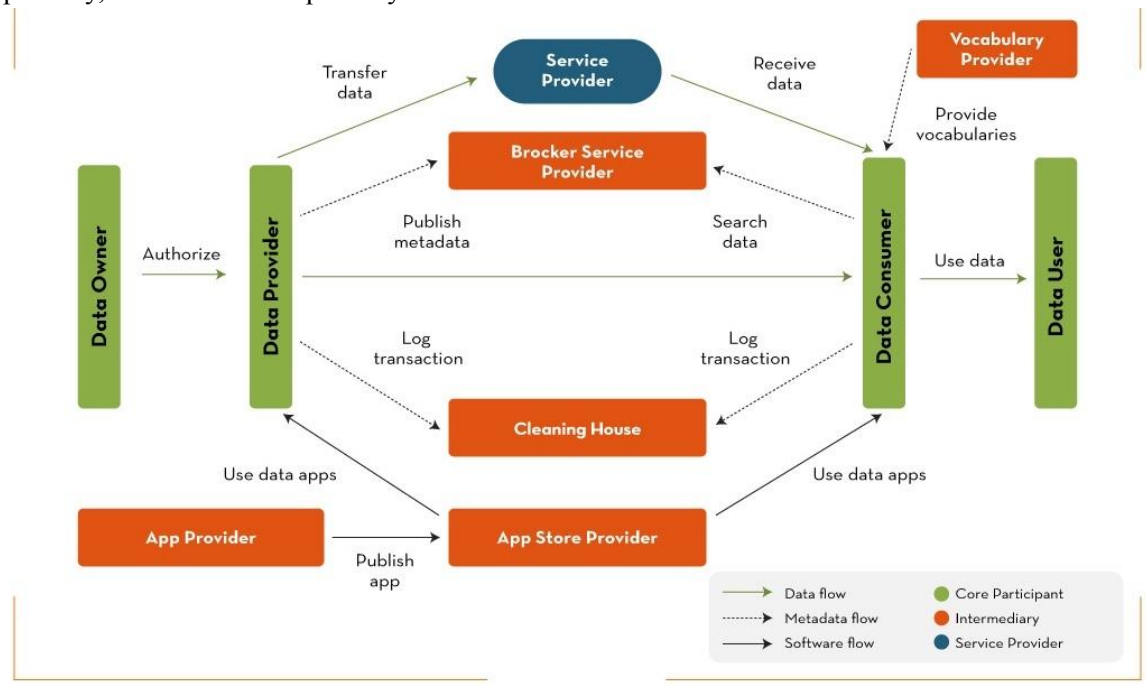

Figure 3: High level architecture of IDS-RAM connectors[3]

The **DataSpaces Business Alliance - DSBA** has proposed a Reference Technology Framework, in their recently released Technical Convergence Discussion Document [DSBA, 2023]. This framework is based on the technical convergence of existing architectures and models and leverages mutual infrastructure and implementation efforts. The goal is to achieve interoperability and portability of solutions across data spaces by harmonizing technological components. The Reference Framework illustrates the concepts of data space connector, data spaces registry and federated services like marketplaces or metadata brokers and how they can be materialized based on open industry standards. To better visualize and understand the details of the descriptions in the paper, the DSBA defined a highly detailed example use case with technical descriptions that can be generalized to other use cases. The use case implements a scenario where a data service provider offers a service on a public

[3]https://datos.gob.es/en/blog/ids-ram-reference-architecture-model-and-its-role-data-spaces

marketplace, so that service consuming parties can acquire access to this offering. An overview of the Building Blocks and WOrkflows of the Reference Architecture is excerpted in the following picture:

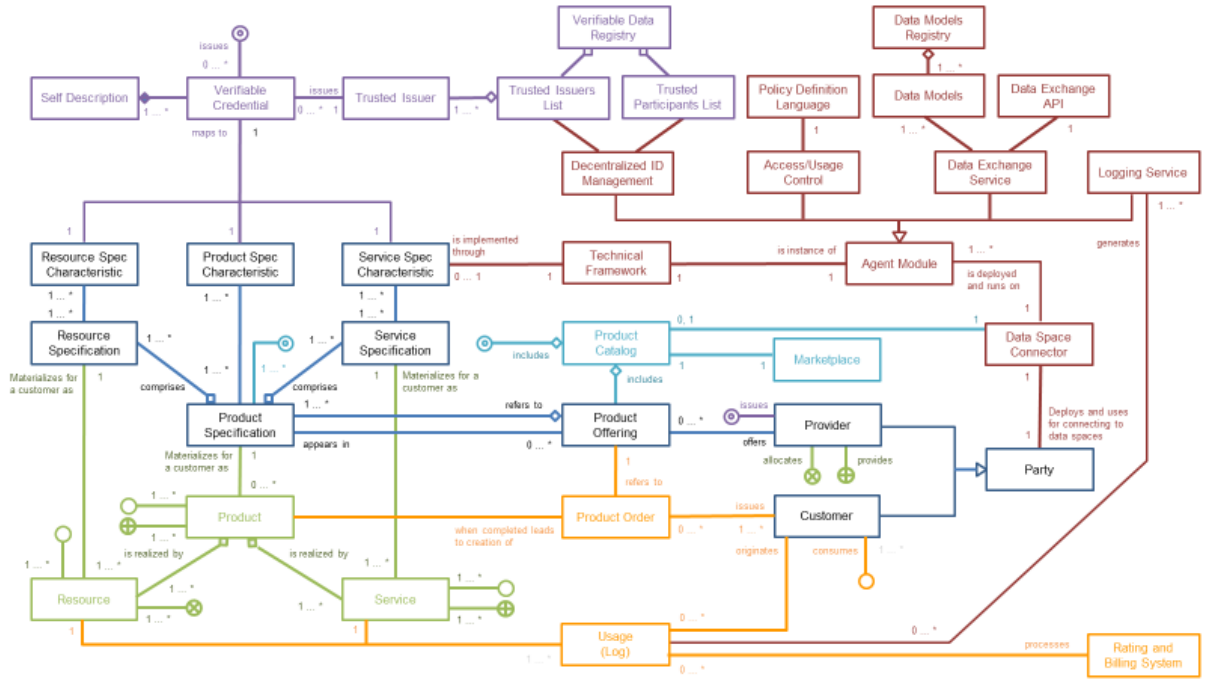

Figure 4: High level architecture of DSBA[4]

The 'Technical Convergence Discussion Document' is an agile paper that will continuously be updated.

All these initiatives rely mostly on european (EOSC and IDS), private calls (Gaia-X) or collaborative effort (DBSA) for the creation of nodes which will be connected to them implementing the standards and policies described above. In the next section we will illustrate some projects and initiatives in the ERA framework.

3. Main Initiatives and Projects

In this section, some of the main initiatives about dataspaces are discussed.

Funded by the European Commission under the Digital Europe Program, the mission of the **Data Spaces Support Centre (DSSC)**[5] is to coordinate all relevant data spaces initiatives in Europe. Among other activities the DSSC defines common requirements and establishes best practices. The DSSC project is part of the European Data Strategy, whose aim is to build a data ecosystem in Europe through the development of common data spaces in strategic economic sectors and domains. The International Data Spaces Association (IDSA)[6] is one of the participants of the DSSC. IDSA is a not-for-profit association representing several industry sectors, with members based all over the world.

The **Europeana Network Association (ENA)**[7] is a community of digital cultural heritage experts with a common goal of enhancing access to Europe's digital cultural heritage through the Europeana

[4]https://datos.gob.es/en/blog/ids-ram-reference-architecture-model-and-its-role-data-spaces
[5] https://dssc.eu/
[6] https://internationaldataspaces.org/
[7] https://pro.europeana.eu/

platform. The platform provides access to digitized cultural heritage assets, such as artworks, books, manuscripts, and photographs. Users can explore and discover diverse cultural treasures from different periods and regions, contributing to the collective understanding and appreciation of Europe's cultural heritage. Europeana aims to play a pivotal role in driving the digital transformation of the cultural heritage sector. Their focus is on developing expertise, tools, and policies to embrace digital advancements and fostering partnerships that encourage innovation. They strive to make cultural heritage more accessible and usable for purposes such as education, research, creativity, and recreation. Europeana's efforts contribute to creating an open, knowledgeable, and creative society. Europeana envisions a future where the cultural heritage sector harnesses the power of digital technology, which in turn leads to a resilient economy, increased employment opportunities, improved well-being, and a strengthened European identity. They actively participate in the common European data space for cultural heritage, a flagship initiative of the European Union that supports the digital transformation of the sector.

We now discuss the main projects related to dataspaces, in particular, the European Union supported the creation and maintenance of dataspaces, as shown by the following projects.

The **IDS (International Data Spaces) Radar**[8] refers to a tool or framework that provides insights and information about the status, development, and trends within the International Data Spaces ecosystem. It offers a comprehensive overview of the key components, technologies, and activities related to IDS. The IDS Radar helps stakeholders in understanding the landscape of IDS, including its architecture, standards, and use cases. It showcases the various organizations, projects, and initiatives involved in implementing IDS and promotes collaboration and knowledge exchange within the IDS community. By using the IDS Radar, individuals and organizations can stay updated on the latest developments, innovations, and advancements in the field of data spaces. It serves as a valuable resource for decision-making, strategy formulation, and identifying potential partners or opportunities in the IDS ecosystem.

The focus of the **data space for security and law enforcement (DIGITAL-2022-DATA-SEC-LAW-03)** should be on facilitating innovation, not covering data sharing for investigative purposes. The objective is to establish a federated data infrastructure and develop a data governance model. Tasks include developing a reference architecture, defining data standards, and establishing criteria for certifications and product quality. Data should be generated, collected, annotated, and made interoperable for testing AI algorithms and security research purposes. Monitoring processes should ensure data quality and validation of results, with a focus on technical standards and unbiased content. Trust mechanisms and data services must ensure security, privacy, and access rights. Efficiency and interoperability within the domain should be considered for data collection alternatives. Fundamental rights challenges should be addressed, including bias mitigation, non-discrimination mechanisms, and enhanced data quality. Compliance with EU legal frameworks on data processing for police purposes and GDPR is crucial. Coordination with relevant projects and adherence to common standards, including the European Data Spaces Technical Framework, are required.

The objective of the **data space for digital communities (DIGITAL-2022-CLOUD-AI-03)** is to pilot and apply the principles of the data space for smart communities on a large scale, connecting data from various domains. The data space will be controlled by public data holders, using open standard tools and supported by a common middleware platform. Funding will support a consortium of stakeholders to foster innovation among EU cities and communities, complying with sector legislation. Pilots will generate a common understanding of progress towards the Green transition and ensure compatibility with the principles of the New European Bauhaus. Cascading grants will support pilots that combine data from areas such as traffic management, climate change adaptation, energy management, and pollution reduction. Pilots should leverage existing infrastructure and make AI services available through trusted application catalogs and marketplaces. Ethical AI solutions, AI

[8] https://internationaldataspaces.org/adopt/data-space-radar/

algorithm registries, and compliance rules should be established at the local level. Links will be established with Horizon Europe missions working with communities and cities for testing and upscaling the data space. Partnership with the Data Spaces Support Centre will ensure alignment with the Smart Middleware Platform and data space ecosystem. The collaboration will focus on reference architecture, standards, interoperability, data governance models, and business strategies.

Data space for mobility (DIGITAL-2022-CLOUD-AI-03) aims to contribute to the development of the common European mobility data space in compliance with EU legislation, creating a technical infrastructure and governance mechanisms for cross-border access to key mobility data resources. The project will align with existing and upcoming mobility and transport initiatives to become part of the European data and cloud services infrastructure. Data related to sustainable urban mobility indicators and traffic/travel information will be made available in a machine-readable format for innovative services and policymaking. The project will support sustainable urban mobility planning by providing data on indicators such as greenhouse gas emissions, congestion, and travel times. It will also provide traffic and travel information at the urban level, following ITS Directive regulations on real-time traffic and multimodal travel information. Projects should have a clear European dimension and involve cities or regions from at least three eligible countries sharing common objectives. Compliance with the European Data Spaces Technical Framework is required, and coordination with other projects and the Data Spaces Support Centre is necessary for interoperability and integration of standards. The smart middleware platform and tools can be utilized, and data accessibility through National Access Points under the ITS Directive is encouraged. The project will ensure interoperability, portability, and integration across infrastructure, applications, and data.

PrepDSpace4Mobility[9] is a 12-month project focused on establishing a secure and controlled method of pooling and sharing mobility data across Europe. It aims to contribute to the development of a common European mobility data space by analyzing existing data ecosystems, identifying gaps and overlaps, and proposing common building blocks and governance frameworks. The project team consists of experts from both private and public mobility sectors, with expertise in mobility, economics, and digital technologies. They aim to facilitate a new era of mobility data sharing in Europe, based on trust, interoperability, and data sovereignty. PrepDSpace4Mobility is a crucial component for the future implementation of a unified market for mobility data. The key objectives of the project include identifying European data ecosystems in the mobility and logistics sector and creating a comprehensive catalog that summarizes relevant data ecosystems and provides information about the type and quality of data.

ENERSHARE[10] develops a Data-Driven Reference Architecture for the energy sector, aligning with FIWARE, IDSA, and GAIA-X standards. It establishes a marketplace using Blockchain and Smart Contracts to enhance trust among ecosystem participants and ensure data security. Additionally, it enables a compensation system, allowing the exchange of energy-related assets and resources (such as datasets, algorithms, and models) for energy assets and services (including heating system maintenance and surplus energy transfer). Engineering leads the project consortium of 30 partners and plays a crucial role in the development of the Energy Data Space that emerges from the project.

D4Science[11] [Assante, 2019] promotes Open Science through implementing innovative Data Infrastructure services which are used by several communities in a common and integrated environment. It faces the open challenges described in section 1.2 and is a pilot for the EOSC initiative in order to publish and share the services of its communities. The platform itself is based on the gCube framework which is specifically conceived to deal with data-intensive science (see also e-Science). In such a domain space, (potentially large-scale) datasets come in all forms and shapes from huge international experiments to cross-laboratory, single laboratory, or even from a multitude of individual observations. D4Science is a candidate technology to create a standard for European

[9] https://mobilitydataspace-csa.eu/
[10] https://www.eng.it/en/case-studies/enershare-il-dataspace-europeo-sull-energia
[11] https://www.d4science.org/

dataspaces and to create a bridge between them and the EOSC. Example of project/communities following this strategy are: Blue-Cloud[12] and SoBigData[13].

At a non-EU international level, the data space ecosystem is apparently underdeveloped. For example, the Administrative Data Research in UK developed the Local Data Spaces project[14] to help local authorities tackle the Covid-19 pandemic. Nevertheless, their dataspace model does not fit exactly the constraints and technological features described in previous sections of this abstract.

Acknowledgements

The work of M. Atzori has been partially supported by MIUR under the PRIN 2017 project "HOPE" project HOPE "High quality Open data Publishing and Enrichment" (prot. 2017MMJJRE) and project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU).
The work of R. Trasarti has been partially supported by MIUR under the "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" project receiving funding from European Union – NextGenerationEU – National Recovery and Resilience Plan – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021."
We wish to heartfully thank Angeles Tejado, Senior Program Manager at FIWARE, for her precious indications, especially for the DSBA Reference Framework, and for her review and suggestions towards the improvement of the Camera ready version.

References

[Assante, 2019] M. Assante et al. (2019) Enacting open science by D4Science. Future Gener. Comput. Syst. 101: 555-563 10.1016/j.future.2019.05.063

[Curry, 2020] Dataspaces: Fundamentals, Principles, and Techniques. In: Real-time Linked Dataspaces. Springer, Cham. https://doi.org/10.1007/978-3-030-29665-0_3

[Dutkiewicz et al., 2022] Lidia Dutkiewicz, Yuliya Miadzvetskaya, Hosea Ofe, Alan Barnett, Lukas Helminger, Stefanie Lindstaedt et al.: Privacy-Preserving Techniques for Trustworthy Data Sharing: Opportunities and Challenges for Future Research. Pages 319-335 (2022)

[DSBA, 2023] Data Spaces Business Alliance, "Technical Convergence Discussion Document", Version 2.0, 2023.
https://data-spaces-business-alliance.eu/wp-content/uploads/dlm_uploads/Data-Spaces-Business-Alliance-Technical-Convergence-V2.pdf

[EGI, 2021] Dietrich, Mark, & Ferrari, Tiziana. (2021). Governance, Architectures and Business Models for Data and Cloud Federations: the EOSC and GAIA-X Case Studies (1.0). Zenodo. https://doi.org/10.5281/zenodo.5081865

[Franklin et. al., 2005] Michael J. Franklin, Alon Y. Halevy, David Maier: From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. 34(4): 27-33 (2005) https://doi.org/10.1145/1107499.1107502

[12] https://blue-cloud.org/
[13] www.sobigdata.eu
[14] https://www.adruk.org/our-work/browse-all-projects/local-data-spaces-helping-local-authorities-tackle-the-covid-19-pandemic-362/

[Halevy et al., 2006] Alon Halevy, Michael Franklin, David Maier: Principles of dataspace systems. Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '06) (2006) https://doi.org/10.1145/1142351.1142352

[IDS-RAM 2019] International Data Space Association, Reference Architecture Model, v.3.0, April 2019. Available at:https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf

[Jarke, et al., 2022] Matthias Jarke, Christoph Quix: Federated Data Integration in Data Spaces. Designing Data Spaces 2022: 181-194

[Nargesian et al., 2019] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, Patricia C. Arocena: Data Lake Management: Challenges and Opportunities. Proc. VLDB Endow. 12(12): 1986-1989 (2019)

[Scerri et al., 2022] Simon Scerri, Tuomo Tuikka, Irene Lopez de Vallejo, Edward Curry: Common European Data Spaces: Challenges and Opportunities. Pages 337-357 (2022)

[Licia et al., 2021] EOSC Architecture and Interoperability Framework, Licia Florio, Mark Van de Sanden, Diego Scardaci, Michelle Williams, Owen Appleton, Paolo Manghi, Keith Jeffery. Public Report