INESC TEC

June 2025

The Future of Data Spaces: Enabling Sovereign, Interoperable and Trusted Data Ecosystems

Position Paper

— www.inesctec.pt

Executive Summary

The data spaces movement serves as a lighthouse contributing to the FAIR – Findable, Accessible, Interoperable and Reusable data sharing principles in Europe, setting to make data truly accessible for citizens and businesses, while ensuring usage limits and ownership rights.

There are nowadays challenges for data spaces to adapt to the wider deployment and use of data intelligence to generate value, and on how raw data, knowledge and generated data can by shared, contributed to and monetized from.

This document contextualizes the data spaces ecosystem and shapes how it could move toward an ecosystem that is capable to handle large volumes of heterogeneous data, while at the same time keep and upgrade how data is discovered, accessed, queried and integrated into industry use cases. At the same time, it provides direction on how data quality and trustworthiness can be brought in to guide data and value sharing among data space participants to benefit from data intelligence. Moreover, it introduces how the benefits from non-authoritative identity schemes are to be integrated as part of Europe's move towards federated identity mechanisms.

The context and challenges feed the vision and characteristics for a data space connector under development by INESC TEC as to tackle the identified challenges and setting the pace for a data space connector tailored for data intelligence workloads and processes, capable to support future use cases and services from the upcoming AI factory challenges and next generation data value chains.

Finally, the document identifies two domain specific data value-chains, one in the domain of energy and another one in the domain of health and introduces how these domains can benefit from a sovereign and trusted data sharing ecosystem.

1. Context and Challenges

The concept of data spaces originated several years ago, setting the basis for a domain-agnostic perspective for data handling. The DSSC blueprint defines a data space as "*a distributed system defined by a governance framework, that enables trustworthy data transactions among participants, while supporting trust and data sovereignty. A data space is implemented by one or more infrastructures and supports one or more use cases*".

This concept sets a move from the Web 2.0 – where organizations establish silos to acquire and process data, often relying heavily on services provided by cloud service providers with standalone governance rules – to the proposed Web 3.0 concept, where decentralized yet coordinated data governance schemes, allow organizations to retain control of their data while promoting data sharing, value extraction, and economic growth. This is the basis for a Digital *European Single Market*, where data spaces are set as the cornerstone for data decentralization and federation, encouraging data sharing among well-established data producers and consumers. This framework envisions data usage to be controlled regarding who, when and for which purpose data can be used, *i.e.,* ensuring data sovereignty, thus supporting a robust and collaborative ecosystem guided by law, regulations, and data usage directives.

This concept leverages four key features that are embodied in the technologies that support data space deployments, namely: a) **security and privacy**, b) **quality and integrity**, c) **policy and governance**, and d) **Performance and high availability.**

The rollout of a data space is carried out through five main dimensions, extending the previous key features, namely:

- **Technology**: exporting specifications on adopted standards, references, and software components, to achieve interoperability among data space components.
- **Functional**: detailing the technical and governance building blocks that embody the needed technical services, their dependencies, and ultimately the data standards and interoperability frameworks.
- **Operation**: considering use cases, requirements, processes, and activities.
- **Business**: studying the business model, particularly the incentives around data exchange of domain-specific data.
- **Legal**: enclosing and evolving the necessary legal frameworks, organizational arrangements, and contractual instruments.

While the technology and functional dimensions embody challenges in research and innovation, such as interoperability, privacy, digital identity, ; others as the operation and business dimensions appeal for technology transfer into operational setups for industry adoption. As a pivotal unit between academia and industry, INESC TEC's positioning builds from a consolidated research and technological expertise in data-centric topics to industry and domain-specific needs to establish the future data value-chains (e.g., energy or health) whose focus is on value extraction through collaborative learning and incentives. Due to the sheer need for collaborative and/or federated approaches, the use-cases which will deliver the future data value-chains and will deliver services for some AI factories require

overcoming a set of challenges. Finally, the legal dimension serves as a lighthouse for sovereignty dispute on data and are left outside scope.

Serving the new collaborative/federated data value-chains and including services born in the recent AI Factories will require the capacity to deal with large sets of heterogeneous data, departing from distinct organisations and public sources. Equipping the data value-chains with **unified discovery**, **querying** and **feature extraction** mechanisms becomes essential for services to export cross-domain benefits. Moreover, data value-chains should be sustained by a model of **high quality of service**, **trustworthy computation** and near to zero downtime, guiding towards **large-scale and highly available data storage systems**.
As to ensure interoperability among data spaces and thus relieve usage and value generation from siloed data spaces, the **data space security threshold** should evolve in two major dimensions: a) **identity** and b) **policy enforcement.** Both security dimensions should embody the principles of free movement within the EU space while adopting the latest non-authoritative identity schemes and approach the EU identity wallet to take data spaces beyond B2B usage and bridge the gap with citizens, while enforcing and interoperable policy enforcement and data usage limits.

Thus, the next-generation Data Space should address key challenges, namely:

- **Large-scale, highly-available heterogeneous data storage**, enabling a plethora of data sources to be dynamically available through rich querying considering non-functional properties such as privacy levels, data availability (e.g., redundancy) or data completeness (i.e., ensure that data is duly represented in the context of one domain and complete, like a specific smart data model or semantic representation). To accomplish this vision, data space connectors would enable discoverability and, under dynamic contract agreements, negotiate access to data made available by that same data space connector, joining data from several domains and rich multi-format and multi-standard data representations.

- **Efficient and unified data access and querying**, supported by a holistic view of the data capabilities from a domain-specific data space (i.e., data representation, data and format variety, and data set volume), establishing a data federation that can unify querying and merging data from several federations as the result of a query. To accomplish this vision, requirements such as data representations and availability, how they mutate and vary over time, or quality metrics that not only address trustworthiness in terms of what data are needed to implement tree or graph-like, holistic search patterns.

- Quality of service in terms of **scale** or **throughput** should steer the inclusion of **QoS – Quality of Service** as the cornerstone of next-generation control plane. Addressing these challenges shall create an **artificial intelligence (AI)-ready data space**, that can efficiently and effectively support the exchange of **learning models** and distinct data monetization schemes for training and inference.

- **Identity** modules require the adoption of federated identity schemes to decentralize the identity mechanisms and move to the inclusion of end users beyond connectors and organizations. Moreover, the link of ranking or reputation scores for specific domains should also steer the overall confidence to produce and consume data to/from the data space.

- **Data usage policies** should move away from approaches based on the goodwill of participants, e.g., a usage policy may dictate the download of data to be allowed in the scope of a given context but only protects the data producer from the misusage of data by logging data activity through a clearing house component for later decision in a court-of-law. This current limitation could be addressed by installing limits to query capabilities or a pre-analysis of queried results through a data-centric pre-clearing mechanism through an active clearing house.

- **Data sovereignty** enforcement requires orthogonal control over heterogeneous data sources. Effective **policy control** requires **privacy-preserving** capabilities for data at rest and while being queried.

2. High-assurance data space connector

Current data space proposals provide an open data concept based on the sparse exchange and copy of data, enabled through syntactic interoperability smart data models, with introductory support for semantic interoperability approaches. Moreover, at most, processes to negotiate digital contracts that unlock access to digital assets (e.g., documents) are automated on data space connectors.

Given the strategic role that access to operational and real data had, which sponsored the scientific and technological evolution of the Web 2.0 concept, it is critical to:

Large-scale, highly available heterogeneous data storage and quality of service

1. **Provide a link to real-time operational data that can be enabled in data spaces**, enabling a polyglot link to enterprise data systems and lakes.

2. **Support heterogeneous data sources** (e.g., unstructured data, object stores, relational and non-relational databases)

3. **Ensure data availability**, both for operation and to guarantee the persistence of data that requires future proofing (e.g., Digital Product Passport).

4. **Align with industry Code-of-Conduct agreements** providing a technology enabler for domain-centric data exchange.

5. Conceive trust models for data and AI models provided by multiple sources.

Efficient and unified data access and querying

6. **Unify access to data**, spread across heterogeneous sources, and make it more performant and scalable through multiple data querying dimensions.

7. **Effectively and efficiently make granular data items available** for AI model training and inference by means of containerized or low-code applications.

8. **Adopting a democratic use of semantic interoperability data representations** to foster adoption, while supported on automatic link of data concepts and translation from syntactic/semantic/syntactic representations.

9. **Unlock the possibility for semantic reasoning** to fully enable data discovery and link on cross-sector use cases.

Identity, data usage policies and data sovereignty

10. **Enhanced privacy and security** for data storage and processing, a key measure towards trustworthy data spaces and their increased adoption.

11. **Effectively enforce data access and usage policies to enhance data sovereignty**, as current implementations eschew this issue

12. Enhance **self-sovereign identity** (for organizations and final users) schemes to support multiple identity providers and credential trustworthiness based on reputation systems.

13. **Actively verify and control data exchange** (in a non-intrusive way) to **limit or block** transactions that **do not match agreed policies** by means of an active clearing house mechanism.

INESC TEC's expertise in high-assurance software development underpins the current development of its data space connector, designed to include the previous needs to serve new data-value chains. The connector will consolidate and embody high-assurance software research areas, and contribute specifically to the following areas:

1. Heterogeneous data processing and interoperability (data plane) (control plane)
2. Large-scale, highly available data management and storage (data plane)
3. AI workflow optimization (data plane)
4. Data security and privacy (data plane) (control plane)
5. Federated Self-sovereign Identity.
6. QoS for distributed infrastructures/systems (control plane)

The design of the INESC TEC data space connector is depicted in Figure 1. It will:

- Adopt a data plane that establishes a close link with data sources and extends querying capabilities of a data domain (i.e., energy, health, industry, among others) and privacy needs.
- Enable the querying capabilities using mechanisms and technologies that are as domain-agnostic as possible, while supporting an extensible data translation framework that enables interoperability across various data representations and standards through semantic approaches.
- Allow a near-real-time exchange of information with state-of-the-art performance, i.e., supporting high throughput and larger data payloads.
- Adopt a control plane that may enforce QoS and other non-functional properties besides policy control.
- Build on the data space protocol[1] which should become an ISO standard by summer 2025.

[1] https://docs.internationaldataspaces.org/ids-knowledgebase/dataspace-protocol

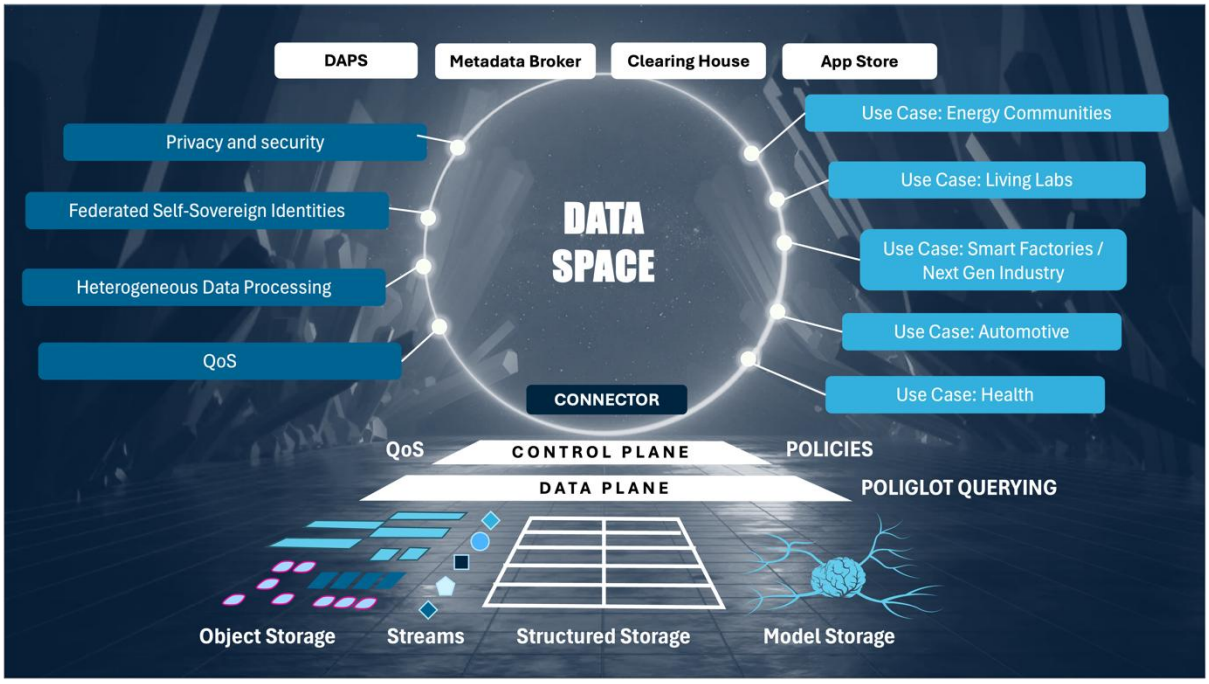

Figure 1 - Vision for INESC TEC's data space connector

The supporting use cases for data spaces, detailed in CEEDs – Blueprint for a common European Energy Data space, focus mostly on the exchange of data for intelligence and AI-based processes supported by the sovereignty guarantees. The processing pipelines will benefit from a connector enabled with support to handle operational data, namely including storage capabilities for columnar, object, streams and AI model data. Moreover, pushing metadata to the data plane may open opportunities for optimizations.

The provision of data services through this data space connector will allow use-cases to benefit from a wide variety of data, equipped with the capability to discover and browse structured, un-structured, timestamp-based data. Most importantly, it will provide a basis to sharing models for inference by including specific connector data planes that will allow the interface with this complex data structures. This, on top of the integrated discovery, high-availability, interoperable and policy capabilities will equip the following two use-cases with trust and curated data sharing from trained models, allowing to set in place and monitor the AI Act in related services.

2.1. Foundational use case: Renewable energy communities and data cooperatives

In the context of Renewable Energy Communities (REC)/ Citizens Energy Community (CEC), data sharing among members and potential members can lead to optimized planning and operation of REC/CEC, unlocking socio-economic benefits and stimulating energy inclusion.

To advance the development and validation of energy data sharing use cases in real-world settings, alongside the design of citizen-centric digital services and technologies at the local community level, INESC TEC partnered with *Cooperativa Eléctrica do Vale d'Este* (CEVE), an energy cooperative acting

as a local DSO and retailer. Together, they deployed a comprehensive range of monitoring equipment including total and individual energy meters, smart plugs, ambient temperature, and humidity sensors across 39 households (first wave of early adopters) with photovoltaic (PV), electric vehicles, heat pumps, electrical water heaters, among others. In this community setting, these tools facilitate a wide array of energy and cross-sector services, fostering innovative consumer and community-centric business models. Furthermore, they ensure critical principles such as privacy, confidentiality, cybersecurity, sovereignty, and granting users full control over their data.

On the top of this data sharing infrastructure, the first use case under demonstration is the instantiation (sizing) of REC and simulation of business models. The goal is to optimize, in the planning phase, the capacity of the distributed energy resources and simulate their operation to estimate an internal reference price to study different business models. For this purpose, combining data from different consumers (i.e., shared by different data owners) is fundamental since it enables the computation of the benefits of belonging to the community and the study of different asset-sharing business models in advance. The data obtained from this REC instantiation can be shared with organizations working to mitigate energy poverty, allowing for a more informed financial assessment of the participation of vulnerable consumers in the REC. For instance, vulnerable consumers can participate in programs where they trade their available space for installing PV panels or battery storage solutions. This can be done through ownership shares in certain assets, reduced energy prices, or by providing non-energy services such as cleaning and inspecting PV panels and maintaining and repairing electrical installations.

During the operation of the REC, another use case is to compute sustainability indicators, such as the traceability of green energy supply within the community or create “happy hours” for community EV charging. Data sharing can also foster the development of circular business models in communities, which are particularly appealing in agricultural settings characterized by seasonal activity. For instance, a surplus of RES can be exchanged for other products/services through barter exchanges, including water, PV panel cleaning, training programs, raw materials, and biofuels. The availability of historical and/or operational data from diverse controllable loads also holds significant value in quantifying their flexibility potential. Access to such data enables consumers to unlock additional revenue streams by exploiting flexibility.

A digital infrastructure for data sharing can also boost energy efficiency actions at the community level, particularly for vulnerable consumers. For instance, it can enable building large datasets by combining small chunks of data (e.g., energy efficiency actions and effects) and increase data volume to enable the application of modern machine learning methods, for instance, to quantify/classify the impact of appliance retrofit or renovation actions in advance. This mitigates investment risks and facilitates the improved design of financial support schemes and policies. The cross-consumer data exchange can also help to build training and education programs for energy efficiency. Another use case is to quantify and predict the energy poverty risk by combining socio-economic data with multiple energy consumption patterns. This information can be used to design targeted assistance programs, such as subsidies, energy efficiency upgrades, or financial guidance.

The scope, data volume, and diversity of use cases within the data sharing community are expanding via involvement in additional EU-funded projects like ENPOWER, HEDGE-IoT, AI-EFFECT, and INSIEME.

These efforts exploit emerging concepts such as data semantic interoperability, aiming to enhance data use across various services that have a tangible impact at the local level.

2.2. Foundational use case: Health research in health data spaces

Data spaces provide an unprecedented opportunity for advances in health research. By breaking free from the constraints of small datasets and data silos, health data spaces produce the opportunity of **inductive epistemological approach to health research**—one that moves from data to theory, uncovering unexpected correlations and long-term effects.

It therefore contributes to enable research into **syndromes**, i.e. conditions that affect more than one health system, e.g. cardiac and dental, cardiac and stomach; or into hard-to-understand conditions, like **chronic pain** (affecting, in 2021, an estimated 20.9% of U.S. adults). Over time it will have **long series** of health, which will enable eliciting long-term causal relationships that were previously difficult to observe. Or unlock research opportunities for **rare conditions**, simply through the law of large numbers. Rare conditions affect 1/2000 people, however, because there are about 7000 types of rare conditions, this means 300 million people globally are affected. Finally, health data spaces present a significant potential to improve the quality of life of traditionally underrepresented conditions and their populations, e.g. genetic conditions more likely to affect certain specific patient groups.

Altogether, this translates into the reduction of prejudicial **bias** in medical research. Today, medical researchers are bound by the proximity of their patients and the diversity of the population in their neighbourhoods, which inevitably means medical outcomes are biased towards the most common populations around research centres.

When we enable AI models to be trained on global, diverse and representative data samples acquired through the sovereign data spaces, we have a pathway to fairer and more efficient medicine. In addition, global data spaces have the potential to uncover patterns in the percentage of people who have side effects of medication.

There is a trend in Health data spaces towards federated learning (FL) of AI models. FL is explored, for example, in EU-funded projects like PHASE IV AI in conjunction with secure multi-party computation techniques to address health data scarcity to train AI models. In this setting, data never leaves the organisation with data custody. This then means the algorithms, tuned by AI researchers, would be sent to organisations with data custody within approved computes, to be trained, explained, evaluated, and scored locally. AI models learnt from different but compatible data sources must be aggregated centrally into larger, more supported models, which reputation must be tracked. In health, copying individual patient data should be avoided, unless explicit granular level permission has been given by the patient case by case. Synthetic data can be generated from real data to preserve privacy by replacing individual values entirely whilst preserving the statistical characteristics of source data. They can however be unfair to locally rare cases.

While preserving identity in healthcare is critical, being able to connect an individual's data across data sources is a requirement by EU Law. Under the new rules, individuals will have **faster and easier access to electronic health data**, regardless of whether they are in their home country or another member state. They will also have **greater control over how that data is used**. EU countries will be required to set up a **digital health authority** to implement the new provisions.

The EHDS will also provide researchers and policymakers with access to specific kinds of **anonymised, secure health data**, enabling them to tap into the vast potential provided by the EU's health data to inform scientific research, develop better treatments, and improve patient care. However, for certain types of research, especially in the case of rare conditions or underrepresented populations, it is critical be to link the same patient's data over time and across diverse sources. This link is broken if data owners individually make their data available in data spaces as an independent experience every time.

Moreover, the new regulation requires all electronic health record (EHR) systems to comply with the specifications of the **European electronic health record exchange format**, ensuring that they are interoperable at EU level.[12]

The scope, data volume, and diversity of use cases within the data sharing community are expanding via involvement in additional EU-funded projects like PHASE IV and AI4SYMMED. These efforts build up on concepts such as data semantic interoperability, federated learning, real-time model availability, aiming to enhance data use across various services that have a tangible impact at the local and global level, which the data spaces concept will contribute to consolidate.

[2] https://www.consilium.europa.eu/en/press/press-releases/2025/01/21/european-health-data-space-council-adopts-new-regulation-improving-cross-border-access-to-eu-health-data/

3. Roadmap

The true adoption of the FAIR principles, namely the *Interoperability* was shaped as part of two EU work programs with consolidates the knowledge, expertise and specific technology being now developed. Figure 2 depicts how did this process started and the foreseen upcoming stages.

Currently, a connector with the described characteristics is underway, including several sub-projects to equip it with the described characteristics for handling large volumes of data, ensuring heterogenous query capabilities and ensure specific data and control planes to guide and monitor the data and model exchange process among data space participants. These developments will deliver the AI ready data space connector, while keeping it interoperable with the ground features of other data space connectors, ensuring compliance with the new Data Space protocol.

As the developed solution matures, engagement will be made with the International Data Spaces Association to seek for certification of the connector.

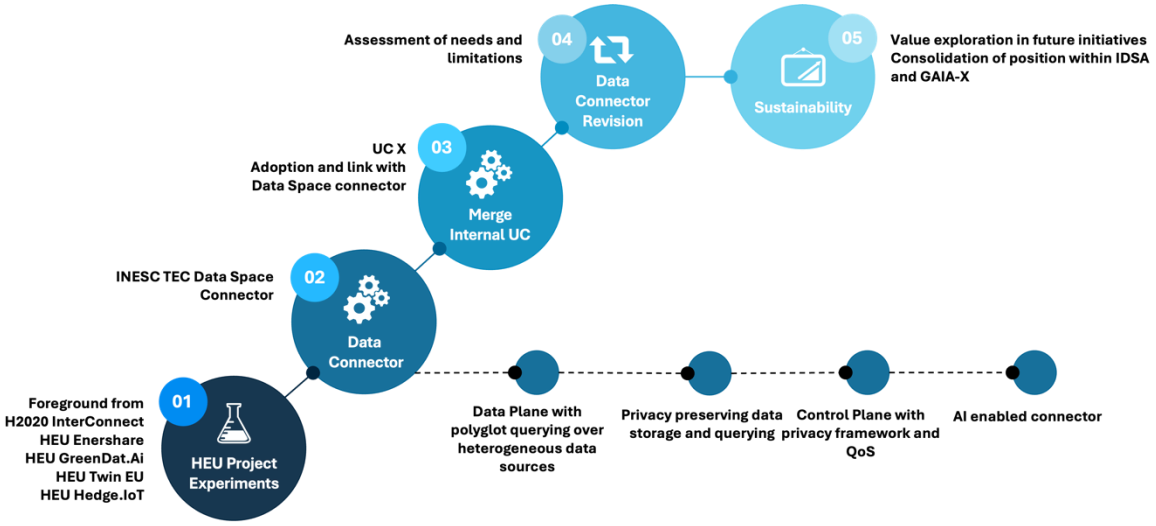

Figure 2 - Roadmap for INESC TEC's data space connector

Along the way, the individual data and control planes will be tested for several TRL targets, considering as much as possible ongoing EU programmed and departing from the context and specific objectives of the figured use cases.

INESC TEC actively participates in a wide range of projects and initiatives related to Dataspaces, spanning sectors from energy systems to personalised healthcare. This breadth of application, combined with a rich variety of tools, methods, and technologies, reflects the organisation's multidisciplinary expertise. INESC TEC's diverse capabilities position it as a strategic and competitive partner in accomplishing Europe's vision of a pluralistic, sovereign, and trustworthy data ecosystem that is ready for AI-driven innovation.

4. Conclusion

INESC TEC's science driven innovation vision congregates a multidisciplinary group of researchers with a wide body of knowledge in computer science, data processing, artificial intelligence and information security, and domain experts in energy, healthcare, robotics or smart manufacturing.

Moreover, INESC TEC actively participates in a wide range of projects and initiatives related to Dataspaces, departing from the design and implementation of core data space's components and spanning sectors from energy systems to personalised healthcare or industry/manufacturing. This breadth of applications, combined with a rich variety of tools, methods, and technologies, reflects the organisation's multidisciplinary expertise. INESC TEC's diverse capabilities position it as a strategic and competitive partner in realising Europe's vision of a pluralistic, sovereign, and trustworthy data ecosystem that is ready for AI-driven innovation.