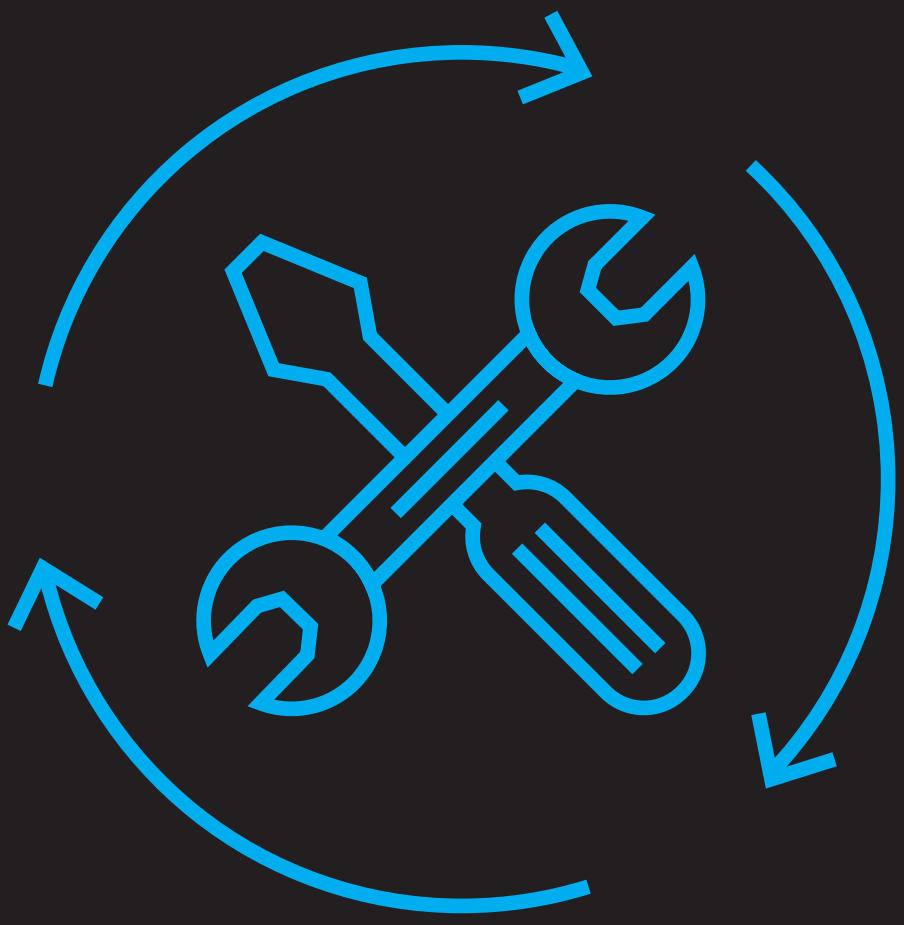
INTERNATIONAL DATA SPACES ASSOCIATION

Position Paper | February 2025

DataBri-X Pilots



- ○ Position Paper of members of the IDS Association
- ○ Position Paper of bodies of the IDS Association
- ● Position Paper of the IDS Association
- ○ White Paper of the IDS Association

DataBri-X

Publisher
International Data Spaces Association
Emil-Figge-Straße 80
44227 Dortmund
Germany

Editor
Olga Galanets, IDSA

Cite as
Galanets, O. et al., DataBri-X Pilots,
International Data Spaces Association,
2025

Authors & Contributors

Anke Losch, Wolters Kluwer
Harry Nakos, Athena Research Center
Javad Chamanara, TIB
Josiane Xavier Parreira, Siemens
Julián Moreno Schneider, DFKI
Konstantinos Oikonomou, NOVA
Marilena Paraskeva, eBOS
Pelayo Fernández Blanco, LSTech
Robert David, Semantic Web Company
Simon Petrac, Nicos
Sotiris Karampatakis, Semantic Web Company
Stelios Sartzetakis, Athena Research Center
Tarmo Luumann, Guardtime

Contributing project

DataBri-X

Funded by the European Union

Funded by the European Union under the Grant Agreement No. 101070069.

This publication is a position paper written by members of the DataBri-X project. IDSA has coordinated the effort that resulted in this paper. The contents of this publication do not necessarily reflect the position or opinion of IDSA, all authors or contributors.

Learn how you can **support IDSA's activities** – visit our website to explore membership options or make a donation under https://internationaldataspaces.org/.

Table of Content

List of Figures

List of Tables

Background and Purpose

The International Data Spaces Association (IDSA) participates in the DataBri-X project as a consortium partner, contributing to the development of practical solutions for data space implementation. The project outcomes address key requirements for data sovereignty, trust, and standardization in European data spaces.

DataBri-X has produced significant technical and methodological contributions to data space implementation, particularly in the telecommunications, energy, and legal sectors. These results include reference implementations, governance frameworks, and tools that demonstrate concrete approaches to data space deployment.

As a thought leader in data spaces, IDSA supports the dissemination of these project results to the broader data space community. The association's established network and expertise provides an effective platform for sharing these implementations with potential adopters and stakeholders.

The Working Groups of IDSA will analyze the project results as part of their ongoing assessment of data space development. This evaluation process helps identify valuable contributions to the field while maintaining IDSA's vendor-neutral position. While IDSA supports the dissemination of all project results, it does not mandate specific implementations or endorse particular technical solutions.

This position paper presents key findings from the DataBri-X project, examining their potential impact on data space adoption and implementation. The analysis focuses on technical feasibility, practical applicability, and alignment with established data space principles.

Profile of the DataBri-X Project

Data Process and Technological Bricks for expanding digital value creation in European Data Spaces (DataBri-X) project started 1st of October 2022. Project consortium comprises 14 partners from 6 European Union members and 1 associated country (United Kingdom), that together form a complete value chain of actors.

The European Data Economy relies on the availability of trustworthy data for innovation, particularly in AI that reflects European values.

Despite the potential of data, the project recognizes that data sharing and interoperability are still in their nascent stages. The diffusion of platforms that facilitate data sharing and the availability of interoperable datasets are critical success factors for driving the European data economy and for facilitating industrial transformation.

To realise a truly cross-border and cross-sectoral data sharing environment, the project's ambition is to create comprehensive Data Spaces and data processing tools which allow for seamless processing of proprietary, personal, and open public data and aims to revolutionize **data sharing ecosystems** by introducing advanced lifecycle practices, innovative tools, and robust governance frameworks. It reimagines how data is managed and shared across industries by introducing innovative strategies that prioritize:

- **Transparency:** Ensuring all stakeholders understand how data is collected, processed, and shared.
- **Efficiency:** Streamlining data sharing processes to reduce complexity and overhead.

- **Collaboration:** Fostering trust and cooperation across sectors and borders.

Project Objectives

The objectives aim:

- to provide tools to support a holistic approach of the data lifecycle in compliance with FAIR principles.
- to build on results of relevant past and current initiatives, data management tools, systems and processes that enable, support and/or automate the creation and maintenance of common ontologies, vocabularies and data models; as well as automated authoring, co-creation, curation, annotation and labelling of data, in view of different later uses (especially AI) of the data.
- to create links with relevant initiatives collecting/using heterogeneous/linguistic data, including AI initiatives (e.g., AI4EU, ELG, or the projects from the H2020 topic ICT-48), and liaise with standardization bodies.
- to provide tools that contribute to the minimization of the energy footprint, be adaptable to different user needs and support and encourage new business models.
- to demonstrate the usability and the value of the tools in diverse use cases.
- to maximise the project's impact and accelerate the adoption and take up of the project in Europe and beyond through wide dissemination, communication, exploitation, commercialization, capacity building and standardisation actions.

Challenges

A critical focus is placed on addressing key challenges in **data management**. Clear frameworks are to be established to define data ownership rights and responsibilities within decentralised data sharing landscapes.

DataBri-X aims to create a secure and sustainable framework, where data sharing is the norm**.** This environment will empower organizations to unlock the full potential of their data while adhering to ethical standards and regulatory requirements.

Mechanisms will be implemented to ensure data provenance and verification, enhancing confidence in data quality.

Robust strategies will safeguard **sensitive data** through effective confidentiality measures and digital rights management, ensuring that digital rights are respected throughout the data lifecycle.

Energy-efficient practices will be integrated into data processing and sharing to minimise environmental impact and promote sustainability.

Decentralised technologies will be explored to enable secure and efficient data sharing, promoting autonomy and privacy without reliance on centralised control.

Implementation Strategy

Implementation strategy includes:

- **Stakeholder Engagement.** Collaborate with industry leaders, policymakers, and data users to co-create solutions that address real-world challenges and ensure widespread adoption.
- **Research and Development.** Conduct research to explore innovative technologies and methodologies that enhance data sharing, focusing on decentralised systems and advanced verification methods.
- **Project Pilots.** Implement pilots to test new management models and tools in real-world scenarios, allowing for iterative improvements based on feedback and outcomes.

Project pilots include:

- NOVA Use Case: The Telecommunication Data space.
- Siemens Use Case: The Energy Data space.
- Wolters Kluwer Use Case: The Legal Data space.

Expected Outcomes

Expected outcomes include:

- A comprehensive management model that redefines data lifecycle practices, fostering a culture of responsible and efficient data sharing.
- Enhanced maturity of data tools and services that are user-friendly, interoperable and secure, empowering organisations to leverage data effectively.
- A framework addressing critical challenges related to data ownership, provenance, confidentiality and energy efficiency, paving the way for a sustainable and equitable data sharing ecosystem.
- Implement IDS-compliant Data Spaces.

1. Introduction

In this chapter, we outline the comprehensive scope of the DataBri-X project, focusing on its alignment with the European Data Spaces initiative.

The scope encompasses the identification and analysis of key components necessary for advancing the European data economy.

Central to our investigation is the re-evaluation of data lifecycle practices. We propose a flexible data governance model that integrates all aspects of data sharing, from discovery to preservation. This model will enhance data usability and discoverability while also emphasising the importance of comprehensive tools and services for data cleaning, aggregation, and quality improvement.

Moreover, the project will address critical issues such as data ownership, provenance, veracity, confidentiality and energy efficiency. These elements will be aligned with the European Union's strong legal framework and policies that advocate for democracy, privacy, protection and equality.

1.1. DataBri-X Conceptual Design

The DataBri-X ecosystem comprises various software tools that provide essential services across different segments of data-intensive projects. These tools are designed to collaborate seamlessly, ensuring smooth data flows and workflows, ultimately enabling users to meet their project requirements effectively.

Figure 1 "DataBri-X Architecture" illustrates the overall project architecture within planning, deployment and workflow execution phases of the project lifecycle focusing on specific aspects of successfully delivering a project.

Planning phase. The process of defining the objectives, scope, resources and timeline of a project. It involves detailed preparation to ensure smooth execution and alignment with goals.

Key components include: **JenPlane Process Designer**, **JenPlane Composer** and **ToolBox**.

JenPlane Process Designer and JenPlane Composer - components enable users to specify project requirements, select appropriate tools, and assemble an efficient workflow for data governance.

The toolbox as a part of the planning stage will also include a Policy Centre that stores customizable policies, ensuring compliance with security and privacy regulations, such as GDPR, while facilitating sustainability and energy-efficient data processing.

The phase's outcome expects a comprehensive project plan that serves as a roadmap for execution.

Deployment phase. The process of implementing the project's deliverables into a live or operational environment. This step ensures that the solution is accessible and usable by its intended audience or stakeholders.

The Tool Deployment includes the workflow builder within running workflows on data, provenance, operation, integration and encapsulation as key dimensions.

The phase's outcome expects a live solution which is operational and ready for use.

Workflow Execution phase. Involves carrying out the predefined processes to achieve specific goals within the project. It refers to the actual implementation of the planned workflows in an organized and systematic manner.

The Execution includes monitoring, orchestration, run time environment and Data Space Connections as follows.

The phase's outcome expects successful completion of processes, leading to the achievement of the project's objectives.

By coordinating these three phases effectively, the project aims to achieve its goals with minimal disruption and maximum efficiency.

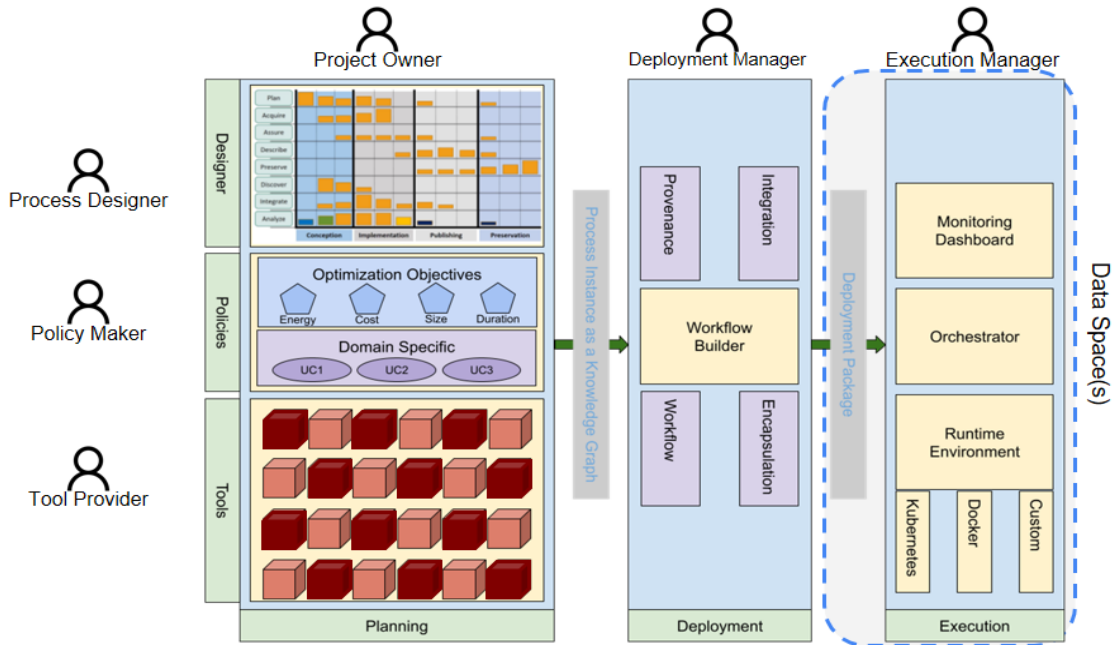


Figure 1 "DataBri-X Architecture"

1.2. JenPlane Governance Process

JenPlane Governance Process offers a model as a flexible, process-driven approach to managing the modern data lifecycle. Unlike traditional linear methods, JenPlane adapts to the evolving needs of data sharing ecosystems by integrating them into a seamless workflow.

JenPlane consists of multiple elements, namely the processes, the designer, the composer, and the builder. At its core, JenPlane empowers users to design their data lifecycle via customization and tailoring of one of the available process templates and then select the most energy-efficient and complementary tools tailored to their specific data management tasks with the help of an LLM-based recommendation engine. This process designer and composer addresses the challenges of tool selection, collaboration, and orchestration, enabling semi-automatic deployment, execution and orchestration of data-driven projects. By creating structured working areas that encompass various disciplines such as planning, data collection, validation, semantic annotation, preservation, discovery and integration, JenPlane facilitates a comprehensive approach to managing data.

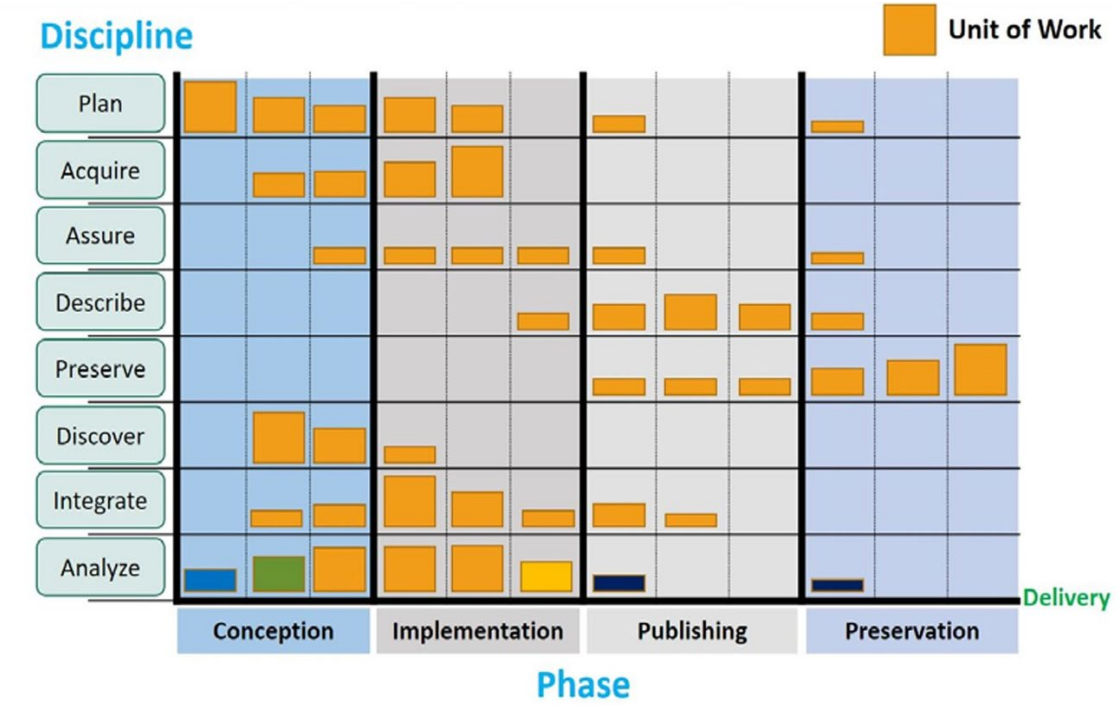


Figure 2 "JenPlane Data Governance Process"

JenPlane Data Governance Process, as shown in Figure 2 above, is a domain specific unique structure, which represents project phases on a two-dimensional axis, ensures flexibility, allows multiple activities to proceed in parallel and enhances adaptability to different data-centric projects.

The process is structured into eight disciplines: **Plan, Acquire, Assure, Describe, Preserve, Discover, Integrate and Analyse,** that define key functional areas and phases that mark the project's lifecycle. Together, they provide a comprehensive framework for decision-making, execution, and oversight.

Each process progresses through distinct phases: **Conception, Implementation, Publishing,** and **Preservation**, with milestones and exit criteria as checkpoints ensuring readiness to proceed. The phases make the workflow more structured. For example, in a conception phase, the user can experiment with different options and possibilities to come up with the most appropriate method, data or even with a tool, and then switching the focus to implementing the actual data analysis pipeline. Upon obtaining proper results from the implementation phase, the focus shifts to prepare them for publishing. After publishing the work in data-intensive projects, there usually comes the need of preservation as a proof of result, as a matter of reproduction, as a matter of archival. Here the focus is on preserving the results via various tools and standards as well as making them available for usage and reproduction.

This governance structure allows users to efficiently identify the necessary data patterns for their tasks and integrate these with other relevant datasets. Following analysis, the results can be published, feeding back into the governance framework for continuous improvement and adaptation.

1.3. DataBri-X Toolbox

The DataBri-X project provides a comprehensive suite of tools – a **Toolbox** designed to facilitate the entire data lifecycle, as an all-in-one solution for managing data in three domain specific Data Spaces: Telecom Data Operator Data Space; Energy Data Space; Legal Data Space.

As shown in Figure 3 "DataBri-X Toolbox" below, each tool utilises a modular architecture which facilitates easy integration into the toolbox, including the JenPlane Composer and other tools as part of the generated workflows. Each tool is associated with metadata describing its functionalities, thereby ensuring interoperability across systems. It includes automated monitoring, data lineage tracking, and provenance features to support real-time insights and comprehensive auditing capabilities. The design is intended to support robust performance management with features like load balancing and autoscaling.

The project aims to improve tools on the TRL level and integrate them into the DataBri-X toolbox that can be configured along the project governance components to be easily deployed in Data Spaces.

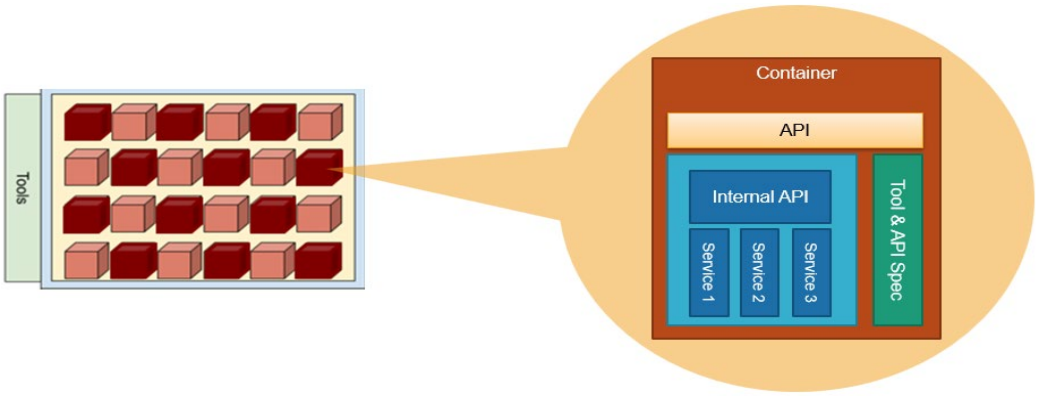


Figure 3 "DataBri-X Toolbox"

In total, 11 DataBri-X partners are providing a set of data tools and services together for different areas of the data lifecycle, as follows: "Acquisition of data", "Assuring of data", "Description and Preservation of data", "Advanced data annotation and analysis", additional tools. Each tool is tailored to ensure that data is captured accurately, validated for reliability, and preserved securely for future use. Let`s have look at them in details:

Acquisition of data is the process of collecting or obtaining data from various sources for analysis, storage and use. Data acquisition includes steps to ensure the data is collected in a format suitable for processing, and it may also involve pre-processing tasks like filtering, transforming, or normalising the data. Effective data acquisition is essential for building accurate datasets that support analysis, research, and decision-making in various fields.

"**Acquisition of data**" include tools such as: TwitHoard[1], LOCI[1], KnowDE[1], SimSearch[1], TripleGeo[1], RWD-E[2], Data Anonymizer[3], NetNous[3], CostNous[3].

[1] Athena Research Centre
[2] GUARDTIME OU
[3] LSTECH ESPANA SL

TwitHoard. Elevating the Twitter experience. It is a dynamic tweet collection tool that adapts to evolving stories, efficiently gathering and summarising diverse tweets, performing sentiment analysis, and offering visual navigation.

LOCI. Capitalising on spatial and temporal data, this tool is an analysis, mining, and visualisation tool. LOCI provides spatial exploration and mining functionalities over points and areas of interest, as well as change detection and seasonality decomposition functionalities over time series data.

KnowDE. Providing data insights with rankings and graphs, KnowDE tackles the data exploration problem by generating efficient knowledge-based and data-based recommendations. The user interacts with KnowDE by selecting alternative keywords and semantically more general or special entities suggested by the system to enhance the user's query.

SimSearch. Implementing top-k similarity search over heterogeneous multi-attribute entity profiles, SimSearch provides versatile ranking capabilities. SimSearch searches and ranks entities consisting of categorical, textual, numerical, spatial, and temporal attributes, spanning multiple local and remote locations.

TripleGeo. Facilitating the interoperability between applications of diverse data format specifications, TripleGeo provides transformation functionalities between semantic web data and typical geospatial data formats. RDF data may be obtained from CSV, JSON, Shapefile, and database data, while the inverse transformation is also supported.

Real World Data Engine (RWD-E). Enabling auditable privacy-preserving multi-party computation. These technologies provide secure, traceable data access and sharing, data reusability, all while respecting legal frameworks and regulations related to security and privacy.

Data Anonymizer. Empowering users to obtain datasets with personal data attributes and anonymize them effortlessly, Dataset Anonymizer ensures privacy by selectively providing specific columns and parameters, offering a secure approach to data handling.

NetNous. Characterising time-series evolution against its baseline and offering insights into trends such as up/down, low, medium, high, assertion and beyond; NetNous provides a comprehensive understanding of data dynamics.

CostNous. Enabling reading from diverse filesystem and data-type technologies, facilitating aggregation, and supporting time-series forecasting, CostNous transforms data exploration.

Assuring data refers to the process of verifying and validating data to ensure it is accurate, consistent, reliable, and meets specific quality standards. This process includes data integrity checks, error detection and correction, and adherence to compliance or regulatory requirements.

"**Assuring of data**" include a tool: Data Integrity[2].

Data Integrity. Ensures the veracity of the data through cryptographically immutable signatures, enabling the verification of the integrity of any-sized data, signing time, and provenance.

Description and Preservation of data involves documenting data thoroughly and ensuring its long-term accessibility and integrity. **Description** includes creating metadata - detailed information about the data's origin, structure, purpose, and any processing applied. This helps others understand and use the data accurately and effectively. **Preservation** focuses

on maintaining the data's usability and preventing degradation or loss over time. This can involve storing data in stable formats, using secure and redundant storage solutions, and regularly backing up the data. Together, these practices support data longevity, reproducibility, and accessibility for future use.

"**Description and Preservation of data**" include tools like: PoolParty Semantic Suite[4], Watermarking[5], Open Research KGraph[6], DataCite Connector[6], RDFox[7].

PoolParty Semantic Suite. A comprehensive software suite offering semantic middleware with components for modelling, loading, and maintaining vocabularies, ontologies, and knowledge graphs. It also includes text analytics, classification, and a recommender engine. With PoolParty, metadata descriptions based on standards are provided for structured and unstructured content.

Watermarking. Advanced watermarking techniques and data types via the Data Provenance API, facilitating data provenance.

Open Research KGraph. Mechanisms, APIs, and tools for publishing data papers and scholarly articles in a semantic and open manner.

DataCite Connector. Integration with DataCite for managing datasets in DataBri-X.

RDFox. A highly optimised knowledge graph and semantic reasoning engine.

Advanced data annotation and analysis refers to sophisticated techniques and methodologies for labelling, interpreting, and deriving insights from data.

"**Advanced data annotation and analysis**" include tools like: PoolParty for Annotation and labelling[4], DataViz[8], LER[9], TIMEX[9], DSR[9].

PoolParty for Annotation and labelling. A semantic platform that uses NLP for data annotation and labelling, transforming data into RDF for easy classification. The RDF annotations can then be used for analysis purposes. PoolParty supports knowledge graph creation, including for legal domains, and ensures that data conforms to required standards.

DataViz. An interactive data visualisation tool that reveals trends and patterns. DataViz's filtering, grouping, and sorting features make data insights accessible, therefore empowering informed decision-making.

Legal Entity Recognition (LER). Tags legal entities in text, enriching documents with RDF-based data for enhanced retrieval. LER is tailored for legal contexts, making document search and linking more efficient.

Temporal Expression Analysis (TIMEX). Recognizes, annotates and normalises temporal expressions like dates, intervals or deadlines in text, making time-based information easily searchable within documents.

Document Structure Recognition (DSR). Identifies and categorises sections within documents, ensuring consistency in structure analysis – ideal for research papers and structured texts.

[4] SEMANTIC WEB COMPANY GMBH
[5] FUNDACION IMDEA NETWORKS
[6] TECHNISCHE INFORMATIONSBIBLIOTHEK
[7] RD SEMANTIC TECHNOLOGIES LIMITED
[8] EBOS TECHNOLOGIES LIMITED
[9] DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH

Additional tools: JenPlane Composer[6], JenPlane Process Designer[6], TRS[6].

JenPlane Composer. Integrated into the Process Designer, the Composer takes a process instance as input and generates a complete ARGO-based workflow with data exchanges, containers, and configurations included. This streamlines the deployment and management of complex workflows on public and private clouds running Kubernetes.

JenPlane Process Designer. Web-based platform that allows users to create accounts, organisations, projects and processes, as well as registering software tools. The designer's main role is to allow users to design their processes, set their optimization objectives and associate their process activities with the operations and services provided by the registered tools. The designer helps the user in many ways, e.g. by guidance on the placement of activities on proper phases and disciplines, as well as by recommending best fitting operations for the automation of the process.

TRS. Provides access to TIB's terminology server and a wide range of terminologies. Supports standardisation and enhances data consistency and discoverability through centralised terminology management.

2. Data Space Integration Architecture

DataBri-X investigates and describes different options to integrate the DataBri-X Toolbox with Data Spaces. The project identifies several requirements that the integration should fulfil and decided for the architecture which best fits these requirements as the following:

- The Data Space integration architecture should be agnostic to specific Data Space standards. Still, we focused on developing an integration aligned with the IDS-RAM. However, the project also aims for GAIA-X integration and generally aims to keep it open how the toolbox integration is implemented and positioned in specific Data Space architectures.
- It aims to integrate the toolbox transparently with Data Spaces and does neither require tools to be aware of specific implementation nor does require changes in their implementation besides the OpenAPI web interface.
- It aims to reuse existing standards, like IDS, and implementations, like the TRUSTS architecture. It builds on standardised existing work and adds to it to fulfil project requirements.

In the following, the project describes the different approaches for Data Space integration and the decision that was made for the DataBri-X project as the following: a (data) pull approach and a (data) push approach.

Pull Integration

The DataBri-X toolbox actively accesses existing Data Spaces via Connector components (pull principle). To do this, the toolbox needs to act as a Data Space participant for every accessed Data Space and run a participating Connector component. Alternatively, each individual tool accesses Data Spaces and implements all necessary requirements for participation.

With this architectural approach, the DataBri-X toolbox and tools can decide and implement their own Data Space access for various Data Space approaches on demand. Individual tools can do so transparently without the need for the toolbox to be aware.

Push Integration

DataBri-X toolbox is accessed via (integrated in) an existing Data Space (push principle) and toolbox workflows are provided via an IDS service provider Connector (currently a DSC implementation). Thereby, the Connector acts as a gateway to the Data Space for the toolbox, which runs behind the Connector. With this architectural approach, the DataBri-X toolbox does not need to know any details about the specific Data Spaces, while making DataBri-X services available to Data Space participants.

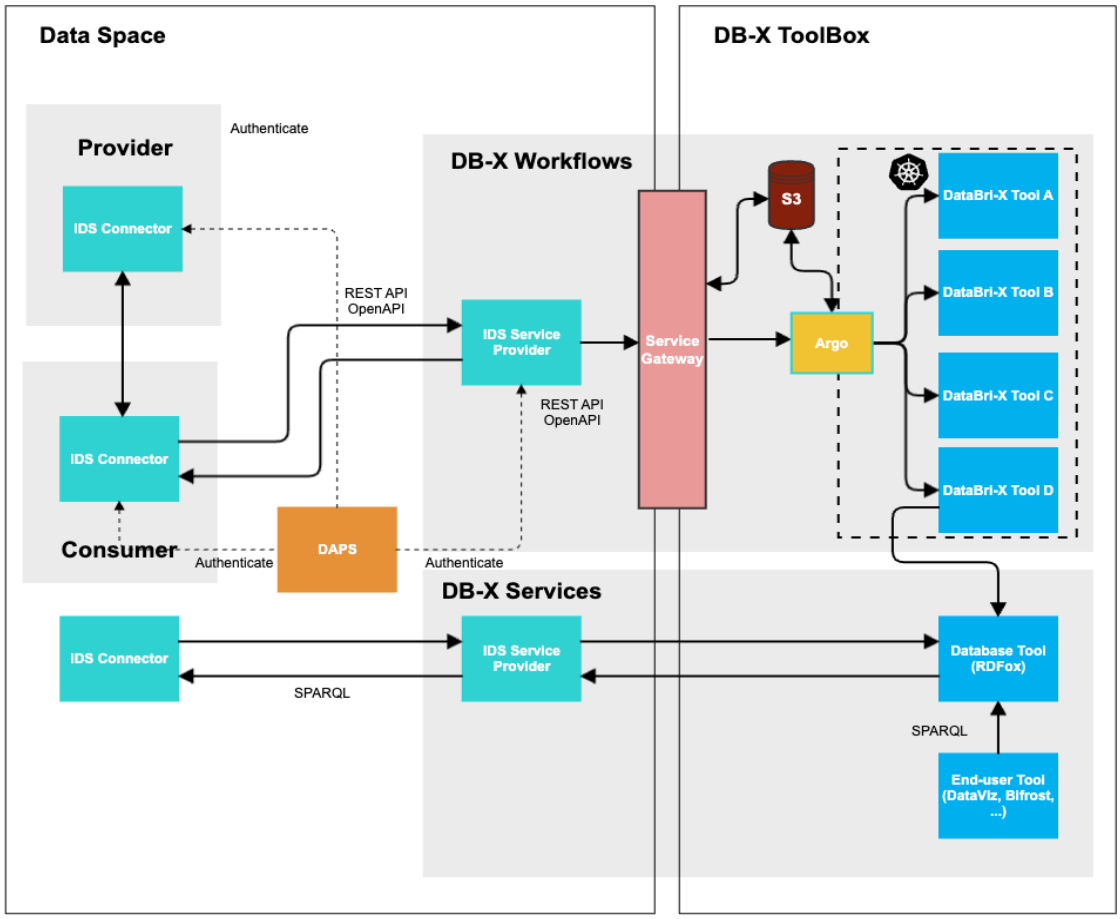


Figure 4 "The DataBri-X Integration Architecture"

Figure 4 shows the proposed architecture and the integration of Workflows (temporal processing tasks) and Services (permanently established) separately exposed via a DataBri-X Service Gateway. IDS Service Provider components make access to the Service Gateway available for the specific Data Space.

Given the requirements identified above, a decision was made to proceed with the push integration for DataBri-X project. The main reason is the transparent access to make the toolbox available via a Service Gateway, which consequently makes it unnecessary to implement specific Data Space standards or run specific Data Space Connectors as part of the toolbox. This most flexible approach was decided to be the best option to make our work adaptable and future-proof. We note that therefore we only show the Data Space integration architecture for the push integration here.

Implementation Report

In the following we describe the developed software components and implementation. We currently base our work on outcomes of the TRUSTS project, which developed an IDS Data Space implementation based on the DSC Connectors.

Artefact/Data Connector Access via HTTP POST

A minimum viable Data Space is composed of a set of connectors and an Identity Provider, typically a DAPS. The connectors can be at the same time-consuming data and providing data,

thereby creating a data exchange environment where participants can share data in a trusted sovereign space. For the sake of simplicity, let us suppose that one of the connectors is the data provider (DP) whilst the other connector is the data consumer (DC). The typical scenario of data exchange in a Data Space is the following. The DP has some data to exchange which the DC wants to use. For this purpose, the DP must describe the dataset using the IDSA data model. Additionally, the DP needs to attach a contract offer to the dataset, bound with a usage policy. Then the DC is able to view this offer by querying the DP connector. After an agreement is achieved between the DP and the DC based on the offered contract, the DC can get access to the actual data. For this purpose, the DSC connector offers several ways to access different data backends. By default, three of them are available, which are local access, access to remote data via external Web APIs and JDBC. Other data access methods can be added by extending the connector implementation, usually by providing Apache Camel components and custom routes:

Direct data access: the dataset is stored on the connector as a blob in the database of the connector. Then the DC can get access to this data by issuing an HTTP GET request on the /artefact/artifactUUid/data endpoint to download the data:

- Relational database access: A database access can be defined as a data backend. The DP provides the configuration of the connection to the database when creating the resource on the connector, specifying the query that will be used to acquire the data. Thus, whenever the DC accesses the data with an HTTP GET request on the DP connector, the DP connector will request the data from the database and send back the data to the DC.
- HTTP Backend: The DP defines an HTTP backend, i.e. an HTTP server, which will provide the data on request. The HTTP backend can be protected with Basic Auth or an API key, which the DP defines on the configuration of the backend. Additionally, any request path or query parameters can be sent to the backend in order to modify the request. However, although the data endpoint can accept any HTTP method, the request to the HTTP backend is always a GET request. Thus, the DC connector can send an HTTP GET request on the data endpoint of the DP connector, which will be executed from the DP connector and the results sent to the DC connector as part of the response.

Based on the default behaviour of the DSC connector, an HTTP POST request is accepted by the data endpoint. However, the actual request to the HTTP data backend is an HTTP GET request. This fact, from the perspective of data exchange is sufficient, since it is supposed that a DC will only want to receive data without sending any data. But in the case of using a service, this is limiting the capabilities. A service provider is performing a service, or a chain of services in the case of workflows, on top of several datasets, as requested by the service consumer. Thus, the service consumer should provide access to these datasets.

As described in the section above, we decided for a push approach on how the DataBri-X Toolbox will be accessed by the participants of a Data Space. That means each Data Space participant that requests to execute a workflow on the toolbox needs to push the data to the toolbox. To achieve this through an HTTP request, the connectors should support the HTTP POST method on the data endpoint. To this end, we extended the capabilities of the DSC connector, to fully support HTTP POST requests on the data endpoint.

In particular, the ArtifactController was extended to support the HTTP POST requests by adding the files request parameter, so that service consumers can post the data needed for

the execution of a workflow on the service provider connector. Moreover, the ArtifactRequest Handler and all the relevant internal messaging modules were updated to ship and retrieve those input files on the connector and forward them on the service provider backend.

The Service Gateway

As shown on the DataBri-X Architecture Diagram, we decided to add an abstraction layer, namely the Service Gateway (SG), between the connector or any other access point, and the actual entry point of the DataBri-X toolbox, the Argo Workflows Server. This solution allows for maximum flexibility, as the SG handles the incoming requests from the different implementations of the connectors and any future versions of those. The objectives of the SG are:

- To authenticate and authorise access to the DataBri-X Toolbox through the Argo Workflows Server. This is provided by a Spring Security component and a connected Keycloak Server.
- To handle workflow input and output artefacts consumed or produced during the execution of a workflow. The SG uses an S3 client to store the artefacts on an S3 storage repository.
- To submit a workflow execution as per client request. This and the following objectives are implemented as gateway routes, preprocessing any incoming request and forwarding the request to the specific Argo Workflow Server HTTP API. Users are not allowed to use any other endpoints. Additionally, the request is validated through this process.
- To provide access to the logs of a workflow execution.
- To provide access to the status of a workflow execution.

In the case of a Data Space, the DataBri-X Toolbox is becoming a member of the Data Space and thus is accessible by other participants, by deploying its own compatible connector (the Service Provider Connector or SPC), registered on the Identity Provider the Data Space uses. The only channel of communication between the participants (and therefore between participants and the DataBri-X toolbox) is through the connectors, utilising either the messaging interfaces in the DSC compliant connectors, or the IDS Protocol in the EDC based connectors. In Figure 5, there is a sequence diagram of the procedures needed to execute a workflow on the DataBri-X toolbox, from the perspective of a Data Space.

As a prerequisite, the workflow definition should be offered through the SPC as an asset on the Data Space, meaning to describe it using the IDS Data Model and attach a contract offer for it with a relevant usage policy. At this stage, the minimum needed usage policy is the Connector-restricted Data Usage policy, which restricts the usage of a specific asset with a specific connector. Thus, only the connector for which the workflow was defined, will be able to make an agreement with the SPC and use the workflow execution infrastructure.

After the workflow asset is defined on the SPC, it is available to the Service Consumer Connector (SCC) to make an agreement and execute a workflow. The SCC should then follow the standard sequence of actions needed to get access to the workflow, as shown on the first sequence in the diagram.

After the agreement is achieved, the SCC can execute the workflow. The definition of the workflow is already defined, but the data to be used is still not bound. The SCC client (usually

a user) can choose which data to use for an execution. The data can be either data assets that the SCC owns or has rights to use in the Data Space or it can be local to the SCC data. The agreement makes the workflow execution artefact available, which is an HTTP endpoint of the local SCC. The SCC client can attach any number of input artefacts on the request that are needed to execute the workflow. The SCC receives this request and follows the second sequence to trigger the workflow execution. At the end of this sequence, the SCC client receives the metadata of the execution, notably the execution ID to further interact with the DataBri-X toolbox.

As the workflow execution is asynchronous, the execution ID is very important, as it is used to get the status of a workflow, the logs and the data output of the workflow.

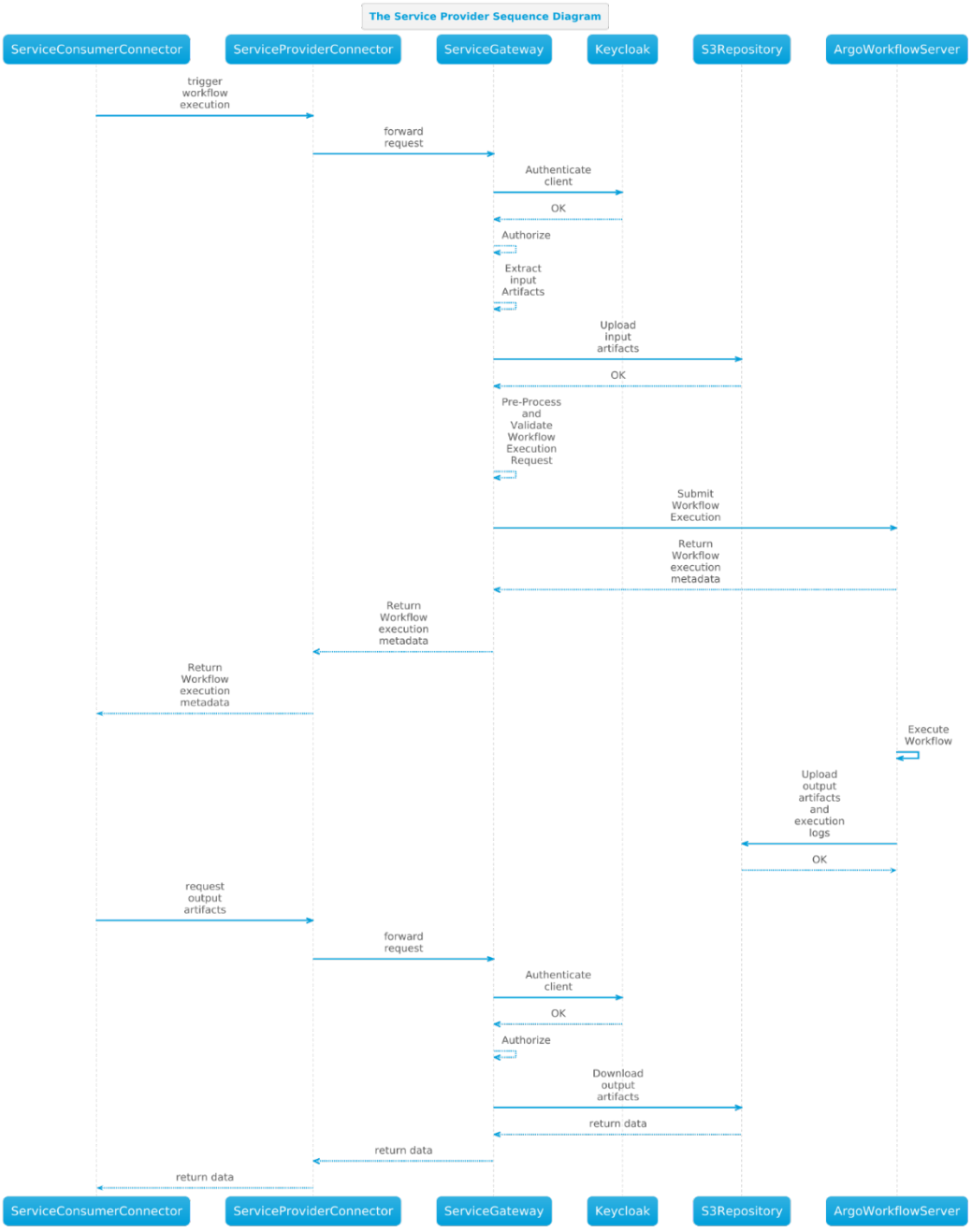


Figure 5 "Sequence Diagram of Interaction with the DataBri-X Toolbox in an IDS Data Space"

The Service Gateway offers the following endpoints:

- POST /submit: Used to trigger a workflow execution.
- GET /status: Used to get the status of a workflow execution. Parameters: workflowname: the workflow execution id as was returned by the submit action.
- GET/logs: Used to get the logs of an execution. Parameters: workflowname: the workflow execution id as was returned by the submit action.
- GET /output: Used to get the outputs of an execution. Parameters: workflowname: the workflow execution id as was returned by the submit action.

The repository of the Service Gateway can be found in DataBri-X GitLab.[10]

Further Adaptations of TRUSTS Components

For the current implementations of the DataBri-X project Use Cases, we use the Data Space components provided by the TRUSTS project, particularly the setup of a TRUSTS corporate node [11]. This deployment offers a fully functional IDS Data Space Connector (commonly known as DSC) and a CKAN deployment. The CKAN instance acts as an interface to the Data Space Connector. In action, the CKAN instance works as any other CKAN deployment is used by several open data portals worldwide. However, this deployment uses an IDS extension, developed in the TRUSTS project, that provides a client to the local DSC connector and all the relevant GUI elements to interact with it. Thus, the deployment offers a user friendly and easy to extend infrastructure in order to deploy a functional Data Space node. The IDP (usually a DAPS server) is not part of this deployment.

In order to make available the workflow execution infrastructure to the nodes, we further developed the ckanext-ids17 extension. In brief, we extended the plugin to offer:

- HTTP endpoints to manage the workflow execution.
- GUI components to manage the workflow execution.
- Workflow execution monitor provides information on the status of a workflow. Through this the user can get access to the workflow execution artefacts, i.e. the status, the logs and the output artefacts.
- Workflow configurator: a GUI component that guides the user in order to create a workflow execution. The component parses the workflow definition to determine the possible input artefacts needed by the workflow and presents the user with a selection menu, to map user owned datasets (either owned or acquired) to workflow input artefacts. Its then creates the request and uses the local connector to submit the workflow.

[10] https://gitlab.com/databri-x/swc/service-gateway

[11] https://github.com/Trusts-eu/trusts-corporate-node

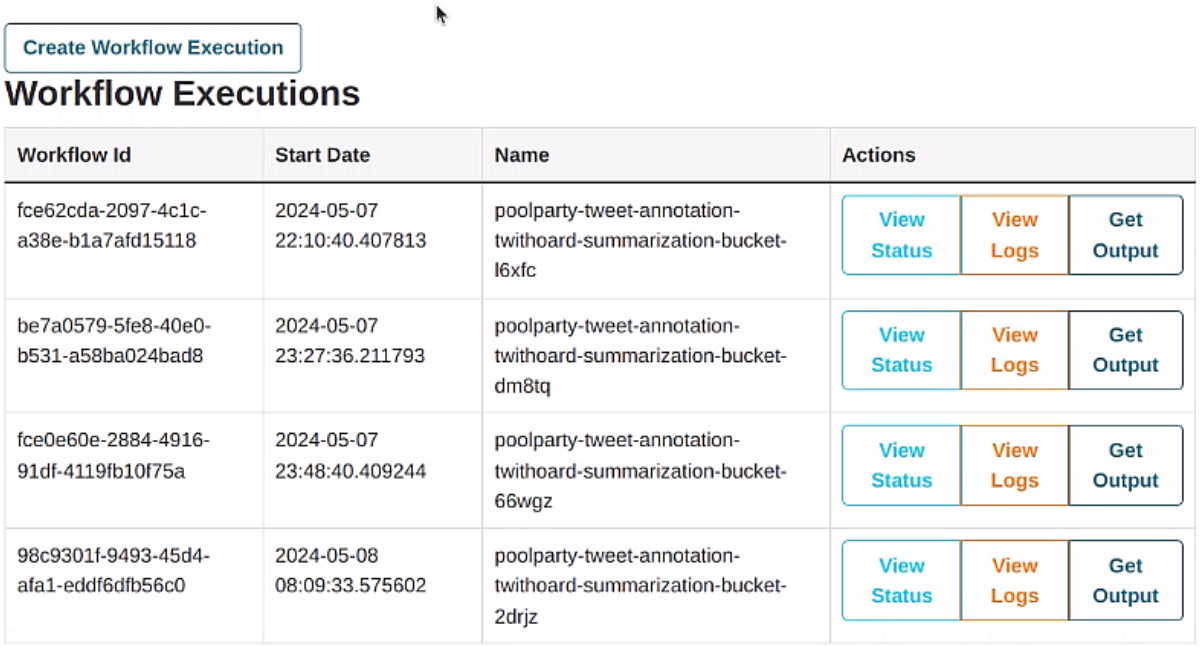


Figure 6 " Workflow Execution Configuration component in CKAN Operational Environment "

The operational environment provides an infrastructure for i) the use case Data Spaces and ii) for developing the Data Space integration. As part of T3.5, NICOS provides a testbed for an IDS Data Space, which acts as a starting point for further developments.

EDC Service Provider

To enable the usage of the DataBri-X Toolbox in Data Spaces that utilise the Eclipse Data Space Connector (EDC), an EDC Connector from Sovity is provided as a Gateway to the DataBri-X Services. The EDC from Sovity is currently developed for the Mobility Data Space (MDS) and supplied as Community Edition for other use cases. A customization of the Sovity EDC is planned to slim down the service to its necessary functionality, but for now the current version of the Sovity EDC is used. There is still further testing needed to ensure coverage of the use case and to identify unwanted components from the Community Edition, but we are very sure that the Sovity EDC will work well with the Service Gateway to ensure complete functionality as a Service Provider.

3. Use Cases and pilots

In this chapter, the DataBri-X Architecture will be described as a Domain-Specific Architecture (DSA) that refers to an architectural design tailored to address the needs and challenges of a specific domain or industry.

The demonstration and work in the DataBri-X project will revolve around three different use cases and their scenarios, trials and revenue streams based on a three-tier digital strategy as follows: follow growing revenues from current services; growing strategic revenues; reducing costs.

The strategy will be executed through three distinct Use Cases, as illustrated in Figure 7:

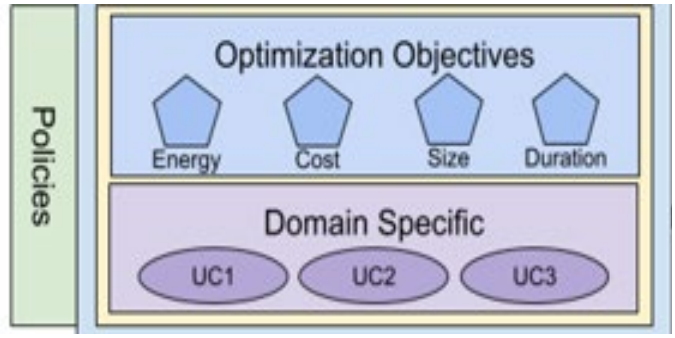


Figure 7 "Domain specific architecture"

Use Case 1: Telecom Data Operator Pilot. This Use Case focuses on leveraging telecommunication data to address key challenges and opportunities in the sector and establishing a Telecommunication Data Space. It includes four scenarios: 1) Scenario 1: Social media brand equity and product reputation analysis. 2) Scenario 2: Call centre chat analysis and interaction improvement. 3) Scenario 3: Infrastructure logs analysis. 4) Scenario 4: User plane records analysis.

Use Case 2: Energy Pilot. This Use Case focuses on advancing energy data management and collaboration within the Energy Data Space. It includes the following scenarios: 1) Scenario 1: Energy Community design and verification. 2) Scenario 2: Onboarding of participants into EC.

Use Case 3: Legal Pilot. This Use Case is cantered on establishing a Legal Data Space to enhance the management, governance, and processing of legal information. It includes the following scenarios: 1) Scenario 1: Create a nucleus for a Legal Data Space. 2) Scenario 2: Add tools and a data governance layer. 3) Scenario 3: Address legal issues with Data Governance and AI.

The following sections dive deeper into each use case, in terms of goals, challenges, inputs and expected outcomes.

3.1 Telecom Data Operator Pilot

Use Case Owner: NOVA, Athens, Greece.

NOVA is a big telecom provider in Greece that belongs to the United Group in Southeastern Europe. Through the DataBri-X project NOVA wants to advance the maturity of data groups, services and promote the data sharing environment in the data sharing ecosystem. Advanced data-driven analytics can be a game-changer for enhancing a business value proposition and improving end-to-end customer experience and that is exactly what NOVA wants to achieve.

NOVA is piloting the Telecom Data Operator Data Space relying on IDSA and Gaia-X standards. The Data Space is designed to harness the full potential of the DataBri-X toolbox and its analytics capabilities. NOVA has established its own Data Space within the DataBri-X project, operating it with a dual role. This dual role as both a provider and consumer within the Data Space, though by different stakeholders, positions NOVA to optimise resource allocation and facilitate seamless data exchanges, driving innovation and value creation.

NOVA is utilising the DataBri-X toolbox and the Data Space architecture to execute multiple scenarios close to the business needs of the company. The following sections dive deeper into each scenario, both in terms of why they were designed and selected and what were the actual outcomes.

Scenario 1: Social media brand equity and product reputation analysis was requested by the marketing department of NOVA, which is intrigued with the capabilities of the demonstrator regarding the analysis of social media data and more specifically the summarization of tweets and discovery of influential users to use for the company's marketing campaign.

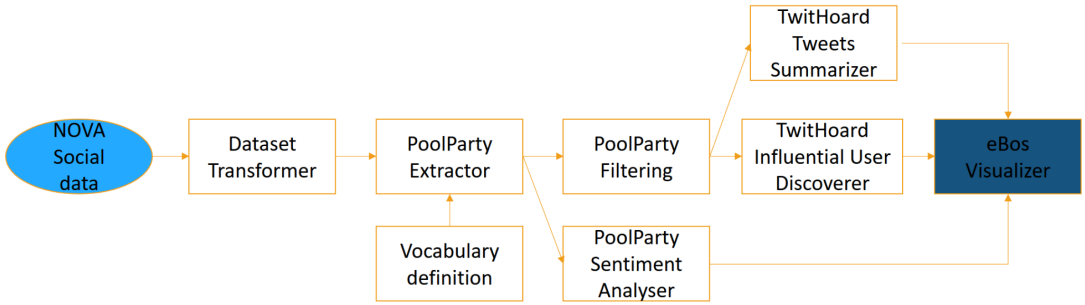


Figure 8 "NOVA Scenario 1"

The datasets that were retrieved from the Data Space and used in this scenario were posts from X (formerly Twitter) in JSON-lines format. The process that was defined in the toolbox contained the following tools:

- **PoolParty:** semantic suite that handles the annotation and enrichment of the dataset. A definition of a keywords vocabulary also allows the filtering of the tweets of the dataset to remove entries that are not relevant to the demonstrator.
- **ELG lexicon and Sentiment Analysis:** A Greek lexicon is imported from the European Language Grid (ELG) to add a value for the sentiment analysis of each tweet.
- **TwitHoard:** The enhanced dataset is used as input to this tool to generate a summarization of the tweets dataset and also to rank influential users based on their impact.
- **Visualizer:** The results are finally fed to this tool to provide a comprehensive visualisation of the performed analysis.

The results of this scenario can be separated into two categories. On the one hand a summarization of the enhanced tweets is produced that contains only those that are relevant to the NOVA company, along with the sentiment score of each tweet. This is invaluable to the marketing department to quickly and efficiently generate tweets analysis and find out the sentiment of users regarding certain NOVA services and the company's public image in general. On the other hand, a ranking of the influence of the twitter users is generated,

providing the marketing department with a list of potential targets for effective and impactful marketing campaigns.

Scenario 2: Call centre chat analysis and interaction improvement was requested by the customer support department to better analyse and utilise feedback from the call centre recordings to properly adjust the call centre agents' strategy and thus increase customer satisfaction in their interactions with the call centre.

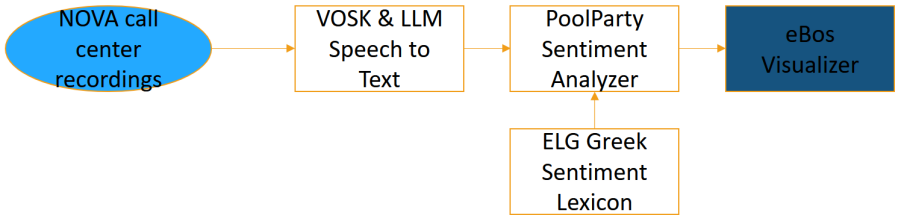


Figure 9 "NOVA Scenario 2"

The datasets that were retrieved from the Data Space and used in this scenario were call centre recordings from the NOVA support in .wav format. The process that was defined in the toolbox contained the following tools:

- **VOSK and LLM:** Audio datasets are processed by a VOSK module to convert them to text. In order to keep the sentiment intact, the results are also processed by a LLM with the support centre context.
- **ELG lexicon and Sentiment Analysis:** A Greek lexicon is imported from ELG to add a value for the sentiment analysis of each word of the recording and an overall score.
- **Visualizer:** The results are finally fed to this tool to provide a comprehensive visualisation of the sentiment analysis of the recordings.

The results of this scenario are a sentiment analysis for each call centre recording that is utilised by the customer support department to monitor and train the call centre agents. Recordings with an overall negative score or with specific highly negative sentences are reviewed and can be utilised for future training of the agents and improvement of the satisfaction of customers in their interaction with the NOVA call centre.

Scenario 3: Infrastructure logs analysis was requested by the network infrastructure department which was very interested in the traffic seasonality analysis and traffic prediction capabilities of the demonstrator and wanted to utilise those to better maintain the NOVA infrastructure as well as reduce the downtime of certain components of it by predicting high traffic moments on those components that might cause a failure.

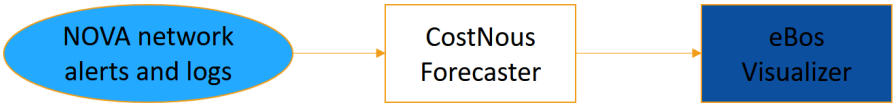


Figure 10 "NOVA Scenario 3"

The datasets that were retrieved from the Data Space and used in this scenario were network metrics from the NOVA infrastructure in .xlsx format. The process that was defined in the toolbox contained the following tools:

- **CostNous Forecaster:** A forecasting engine generates reports for the traffic seasonality of the NOVA infrastructure asset that the dataset corresponds to.
- **Visualizer:** The results are finally put into this tool to provide a comprehensive visualisation of the network seasonality and forecasting.

The results of this scenario are both an overview of the traffic seasonality of the NOVA infrastructure asset that the dataset corresponds to, but also and more importantly a forecasting of the traffic that will burden the asset in the future, based on the patterns of the data, as shown in Figure above with the orange graph lines. This is invaluable for the network infrastructure department as it can raise early alerts for components of the NOVA infrastructure that might become problematic in a specific timeframe in the future, due to increased network traffic.

Scenario 4: User plane records analysis was requested by the marketing department which showed a clear interest in the user clustering capabilities based on the user records, in order to create tailor-made offerings to those users that will appeal to their personal preferences.

Figure 11 "NOVA Scenario 4"

The datasets that were retrieved from the Data Space and used in this scenario were user records of the usage of the NOVA network in .xlsx format. The process that was defined in the toolbox contained the following tools:

- **Anonymizer:** The dataset is marked for sensitive columns that are then anonymized to preserve the identity of the users.
- **CostNous Clustering:** Each NOVA user is assigned to a cluster based on a list of columns, like downlink/uplink throughput used. The columns and number of clusters are fully parameterized.
- **Visualizer:** The results are finally fed to this tool to provide a comprehensive visualisation of the user clusters.

The results of this scenario are clusters of users that utilise the NOVA services in similar ways. The marketing department can use this analysis to provide targeted packages and offers to those users that fit their needs and their utilisation profiles.

Telecom Data Space

The Telecom Data Space, as shown in Figure below, aims to ensure the discovery and utilisation of a dataset via the Data Space, a standardised approach is followed. This approach ensures efficient and monitored access to the data available in the Data Space that were used as part of the scenarios of the demonstrator and consists of the following sequential steps:

- A dataset is inserted in the Data Space by a provider.
- The provider defines the contract that makes that dataset available to anyone interested and registered in the Data Space.

- NOVA, acting as a consumer who is a verified part of the Data Space, discovers the dataset via the metadata broker.
- NOVA accepts the associated Data Space contract for the dataset.
- NOVA gains conditional access to the dataset to be used with the attached workflow for a specific execution instance inside JenPlane.

In this example, NOVA acts both as a provider and consumer, but of course in a shared Data Space those roles can be adopted by a variety of shareholders like other telecom companies, smaller SMEs of the telecom domain without the capabilities of their own R&D department, infrastructure engineers, marketing experts, call centre teams, etc.

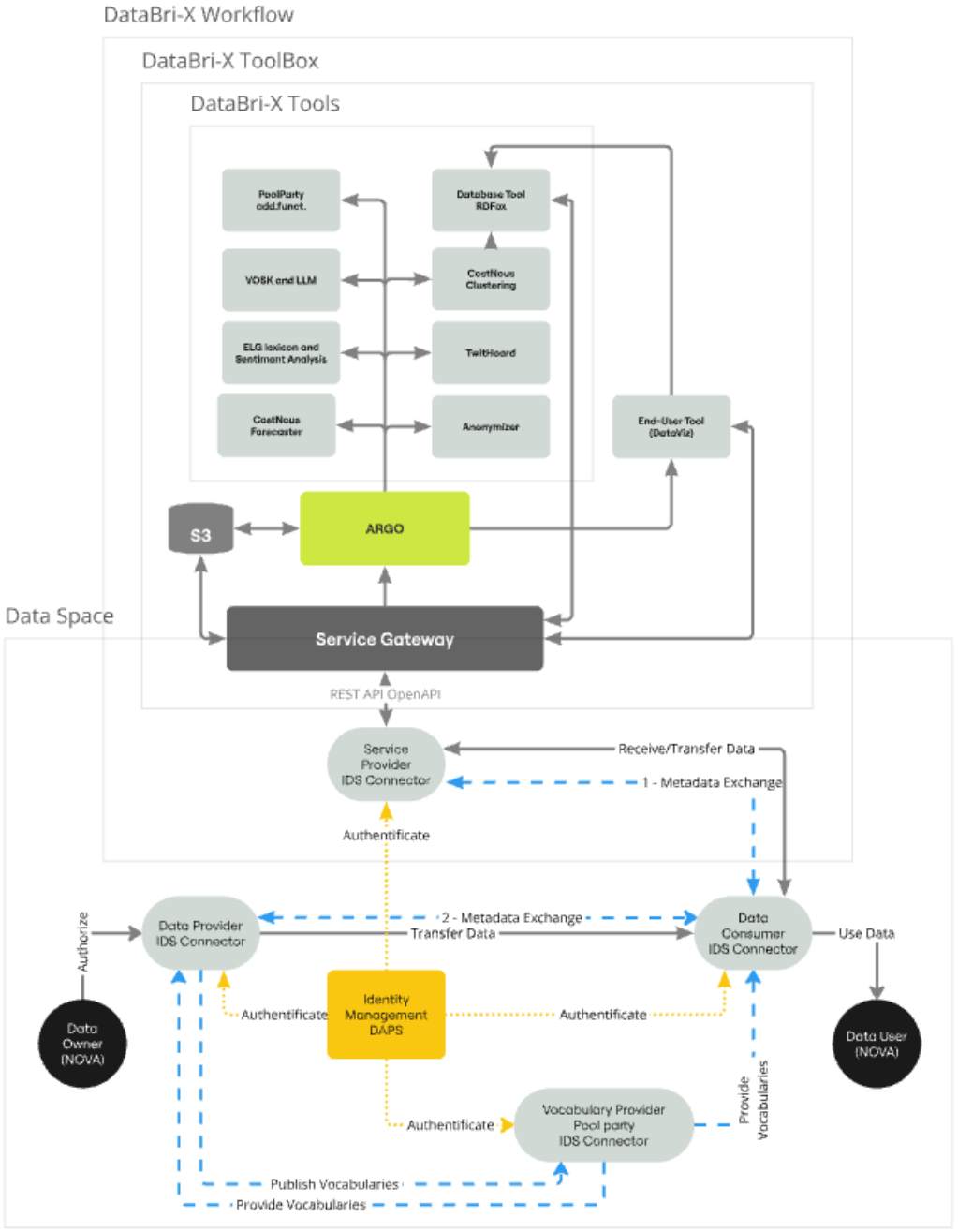


Figure 12 "Telecom Data Operator Data Space"

3.2 Energy Data Space pilot

Use Case Owner: Siemens, Vienna, Austria.

The goal of the Energy Data Space Pilot is to leverage the DataBri-X framework to streamline the simulation efforts for designing and verifying a renewable Energy Community (EC), as well as preparing for the onboarding of its participants in the grid.

ECs are designed to organise collective and citizen-driven energy actions and facilitate the transition towards clean energy in the European energy system. However, their implementation can significantly impact the underlying energy infrastructure. Therefore, before an EC can begin operating, the respective grid operator(s) must establish its technical feasibility to ensure that the community does not jeopardize the security of supply for other users. Simulation can be used to perform this evaluation, but setting up a scenario and preparing the necessary data sources and models for the simulation is a time-consuming task.

Siemens uses the DataBri-X toolbox and the Data Space architecture to execute two scenarios within the EC simulation. The following sections dive deeper in each scenario, both in terms of why they were designed and selected and what were the actual outcomes.

Scenario 1: EC design and verification. This scenario centres on the conception phase and aims to evaluate the economic and technical feasibility of a community during the design phase. It also demonstrates the implications of community participation for stakeholders, particularly the community initiator, and establishes the necessary data sources and services for implementing the simulation. Subsequently, the EC onboarding process can begin. The process/workflow of this scenarios is shown below:

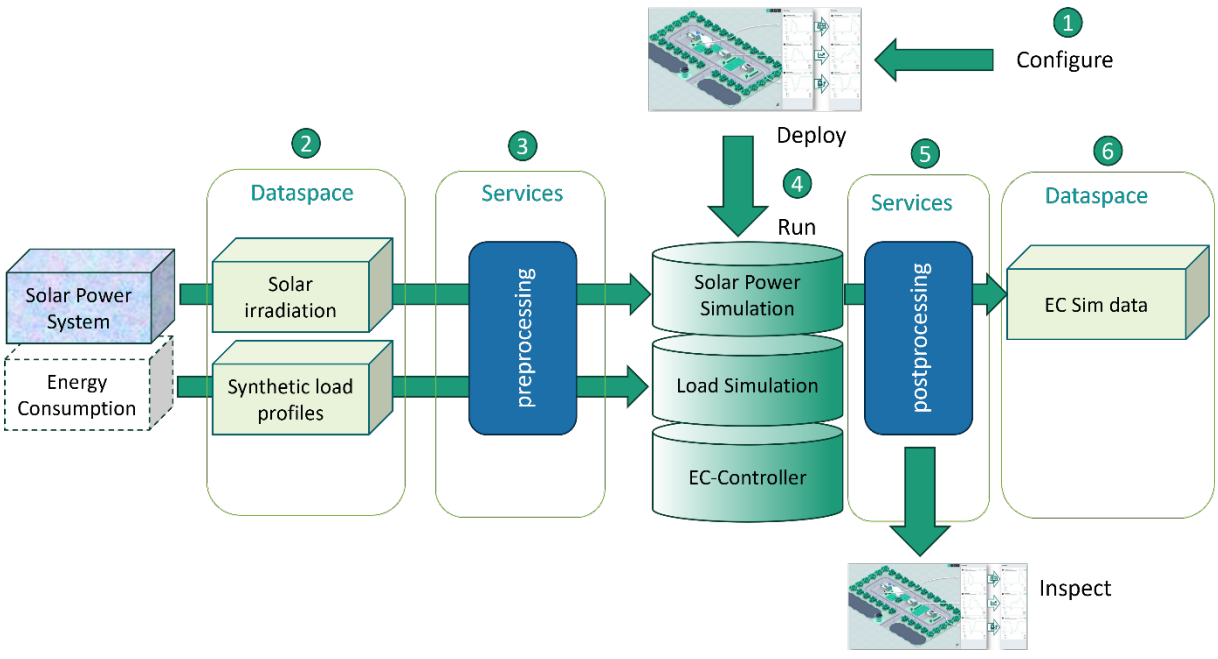


Figure 13 Siemens Scenario 1 "Workflow of the "EC design and verification"

The steps in the above workflow are indicated by the numbers:

1. Manual design of experiment;

2. Identification of suitable simulation data;
3. Preprocessing of simulation data;
4. Energy simulation;
5. Postprocessing of the simulation results;
6. Presenting/publishing simulation results.

The scenario aims for preparing data for the simulation of energy communities. This entails the configuration of the simulation experiment (i.e., selecting a scenario to be simulated). This is a manual step, and it is not within the scope of the project. However, the resulting experimental parameters serve as input for the DataBri-X workflows. Based on the configuration, a set of suitable simulation inputs (time series data and grid topologies) need to be selected from a Data Space in the second step. The data is then pre-processed and made ready for the simulation (step 3). The energy community simulation (step 4) is triggered once all data for the simulation is ready and validated. After a successful simulation, the simulation results are then post processed (step 5) and prepared for (manual) presentation. In addition, results are also published back into the Data Space (step 6).

In terms of DataBri-X tools, this scenario includes:

- **RDFox.** RDFox is used for preprocessing time-series data: using specified rules, it enables the identification and resolution of gaps within the data. Furthermore, it ensures the consistency of time representation throughout the time-series datasets by aligning it with the expected input from the simulation.
- **PoolParty.** PoolParty is used in the use case to check consistency in the EC topology, by specifying and validating SHACL[12] shapes over the Knowledge Graph representation of the EC topology.
- **BIFROST.** BIFROST is used to simulate energy scenarios and extended for being started and stopped programmatically as well as being parameterized with external time series (which were pre-processed via RDFox).
- **IMDEA watermarking.** Watermarking will be integrated to mark generated data before being added to the Data Space, in order to check data integrity.
- **GT Basic KSI.** Similar to watermarking, KSI Data Integrity will be applied to the output data, before sharing it with external partners (via the Data Space).

Scenario 2: Onboarding of participant into EC. This scenario is planned as a consecutive scenario following the previous one. It will utilise the DataBri-X tools of "EC design & verification" and extend them with a hardware in the loop setup, where hardware (an existing simplified building controller) is RESTfully coupled with the energy simulation and can then be tested in the context of the energy scenario, thus easing future deployment, parameterization and integration. The implementation of this scenario is planned for the last year of the project.

Energy Data Space

The Energy Data Space pilot aims at enabling data sharing/discoverability. To this end the scenarios implement a Data Space, as shown in Figure 14, which is used within the use case

[12] https://www.w3.org/TR/shacl/

to: find relevant data sources to run the EC simulation, and to share the data related to the simulation results.

This was achieved by using the Data Space connectors which were developed in the context of the TRUSTS project[13]. A migration[14] is planned.

The Energy Data Space, as shown in Figure 14, consisted of the following sequential steps:

- A dataset is inserted in the Data Space by a provider.
- The provider defines the contract that makes that dataset available to anyone interested and registered in the Data Space.
- Siemens, acting as a consumer who is a verified part of the Data Space, discovers the dataset via the metadata broker.
- Siemens accepts the associated Data Space contract for the dataset.
- Siemens gains conditional access to the dataset to be used with the attached workflow for a specific execution instance inside JenPlane.

In this example, Siemens acts both as a provider and consumer.

[13] https://www.trusts-data.eu/

[14] https://projects.eclipse.org/projects/technology.edc

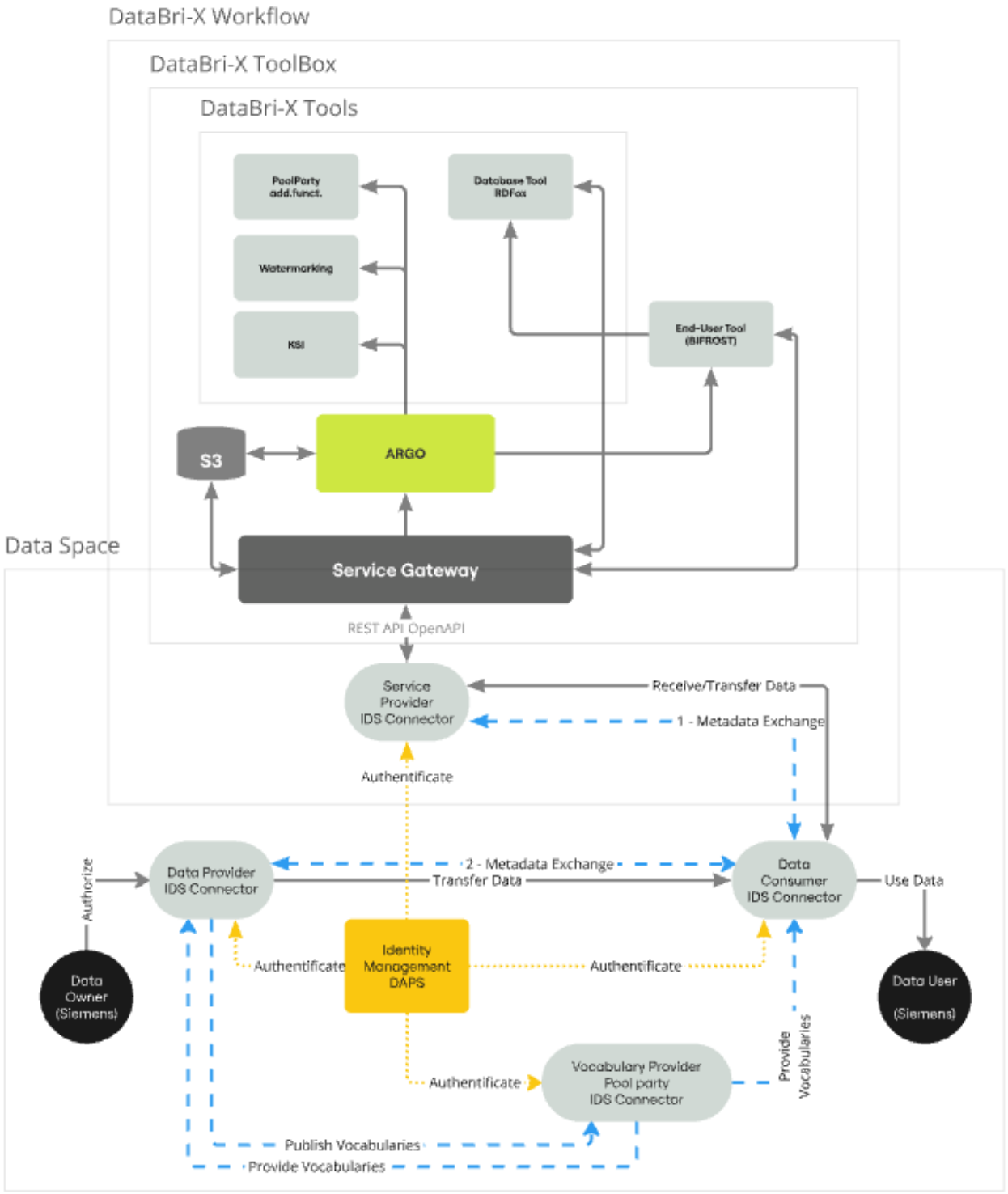


Figure 14 "The Energy Data Space"

3.3 Legal Data Space pilot

Use Case Owner: Wolters Kluwer, Hürth, Germany.

The Legal Data Space pilot focused on promoting ethical standards, ensuring legal compliance, and fostering unbiased data practices within the legal domain. This initiative aims to establish a Core Legal Knowledge Graph, which will serve as the foundation for developing innovative smart services and tools that leverage legal documents and various other sources of legal information.

Key Objectives:

- **Promotion of Ethical Standards**. The pilot committed to upholding the highest ethical standards in data management and utilisation, ensuring that our practices align with industry best practices and foster trust among stakeholders.
- **Legal Compliance**. The pilot will prioritise compliance with relevant legal frameworks, ensuring that all data usage adheres to applicable laws and regulations, safeguarding the integrity of our operations.
- **Unbiased Data Practices**. The pilot aims to cultivate unbiased data environments by implementing methodologies that identify and mitigate biases in data collection and analysis, fostering fairness and equity in legal outcomes.
- **Creation of a Core Legal Knowledge Graph**. By synthesising information from legal documents and other pertinent sources, the pilot will build a knowledge graph. This resource will enable the development of smart services and tools, enhancing access to legal information and facilitating informed decision-making.

The following sections dive deeper in each scenario, both in terms of why they were designed and selected and what were the actual outcomes.

Scenario 1 aims to create a nucleus for a Legal Data Space in order to provide a prototypical platform to store and process legal information.

Table 1 "Wolters Kluwer Scenario 1"

Challenges	Input	Expected results
• Currently no European Legal Data Space exists. • There are quite some open data sources available, but no platform and no coherent governance model or specific tools for applying legal data is in place.	• WKD data. • Open sources data.	The pilot will build a prototypical European Legal Data Space platform, combining DataBri-X tools and methodologies, data from WKD and from open sources and tools from DataBri-X partners and beyond. This prototypical nucleus will be able to process the specifics of legal information.

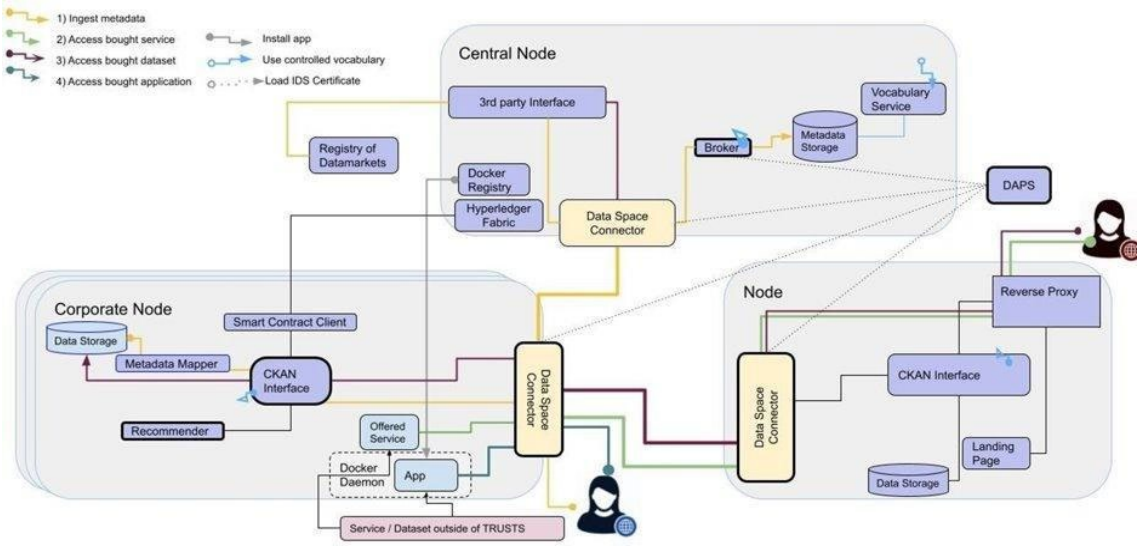


Figure 15 "Wolters Kluwer Scenario 1"

Scenario 2 aims to add tools and a data governance layer to this nucleus in order to be able to solve real world data problems with legal data.

Based on scenario 1, the pilot wants to apply tools and services on legal data, identify and evaluate what specific challenges come with the domain at hand. Legal data has a very specific language, specific style, differences across countries and with that also languages. Therefore, most available analytics technologies do not work well enough with legal data to meet the requirements of lawyers for quality, certainty, and direct usability of the results.

The aim is to empower stakeholders and users with the ability to perform rudimentary data analytics operations, while recognizing the specifics of the legal domain, and integrating various tools and services to easily build processing chains for more value-added outcomes.

There are two main tasks:

1. Data enrichment: In Data Enrichment, the pilot focuses on the tool capabilities and applies as much as possible (quantitative approach).
2. Data analytics: In Data Analytics the pilot focuses on a specific document similarity use case.

Processes need to work independent of a specific data format and in several languages (quality of results can be different).

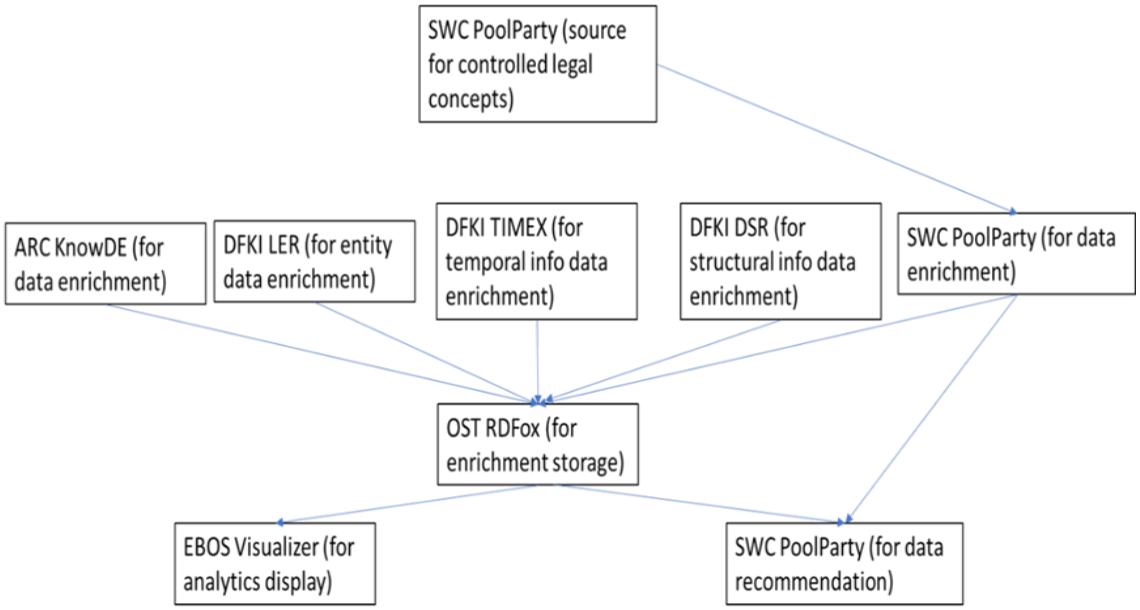


Figure 16 "Wolters Kluwer Scenario 2"

Scenario 3 aims to check on legal issues with Data Governance and AI and provide a forum for discussions and solutioning approaches.

In the context of the Legal Data Space, the core question is why there is no uniform legal Data Space despite its acknowledged importance.

We have a need; we also have data and there could be business around it. Yet we miss several key components.

- Technical standards: we do not have a framework for processing legal data (content standards, data formats etc.)
- Lack of source identification: There is no answer to the question of a qualified indication of origin/source to address questions about data trustworthiness.
- And most importantly the absence of European responses: There are no European responses to the core legal questions that all the mentioned private, scientific, and business users equally ask when placing (their own) data into a Data Space. These questions concern the legal framework conditions, affected rights (copyrights, data protection rights, usage rights, European directives), ownership of data if modified in the Data Space (e.g., by AI applications), ownership of AI trained with licensed data, and the possibility of use or transfer of usage rights based on contracts related to AI applications.

Without answers to these legal frameworks, potential data providers will not make their existing holdings available on a larger scale.

The pilot will provide the basic infrastructure to set up a Data Space environment with required functional capabilities like upload, search, download, check, etc. of data as well as non-functional capabilities like security, performance, transparency, etc. (mainly scenario 1).

This basic infrastructure will be provided by the TRUSTS[15] platform implementation, the DataBri-X governance layer and the provision of a safe cloud environment with access control etc. provided by WKD.

On top of this basic infrastructure, the pilot will run the DataBri-X tools that work on legal data. These tools can be separated into three categories, whereas some tools can be a member of more than one category. Here is the core list of these assignments:

- Basic tools that enable data storage, transfer, workflow realisation as well as quality control and data characterization: RDFox from OST and PoolParty Semantic Suite from SWC.
- Legal application tools that are dedicated to solving business problems of legal professionals, starting from executing efficient legal research, which includes having a critical mass of relevant legal data in one place: mainly all tools from DFKI like LER, TIMEX and DSR, but also KNOWDE from ARC or Visualizer from EBOS.
- Adjacent tools like Basic KSI from GT that enable important, yet self-contained capabilities especially in the legal domain.
- To show real added value, WKD will - based on existing customer needs - define and implement a limited number of workflows within JenPlane and will let users execute these workflows with WKD data or other data that is uploaded and provided by other users and organisations.

Legal Data Space

For law firms and legal departments, as well as for public administrations, universities, and companies that see themselves as service providers for the legal community, the legal framework of Data Spaces plays a crucial role. It is essential to comply with legal requirements and to make them transparent so that the target audience becomes users of such legal Data Spaces. Therefore, our focus is less on questions related to data exchange in general and more on the specific needs of the target audience, such as in relation to the use of AI tools. Additionally, the tools used must meet the particular requirements of legal data. The nucleus we create is intended to evolve over time into a true legal Data Space in the sense of IDSA.

The Legal Data Space, as shown in Figure 17, consisted of the following sequential steps:

- A dataset is inserted in the Data Space by a provider.
- The provider defines the contract that makes that dataset available to anyone interested and registered in the Data Space.
- Wolters Kluwer, acting as a consumer who is a verified part of the Data Space, discovers the dataset via the metadata broker.
- Wolters Kluwer accepts the associated Data Space contract for the dataset.
- Wolters Kluwer gains conditional access to the dataset to be used with the attached workflow for a specific execution instance inside JenPlane.

In this example, Wolters Kluwer acts both as a provider and consumer.

[15] https://www.trusts-data.eu/

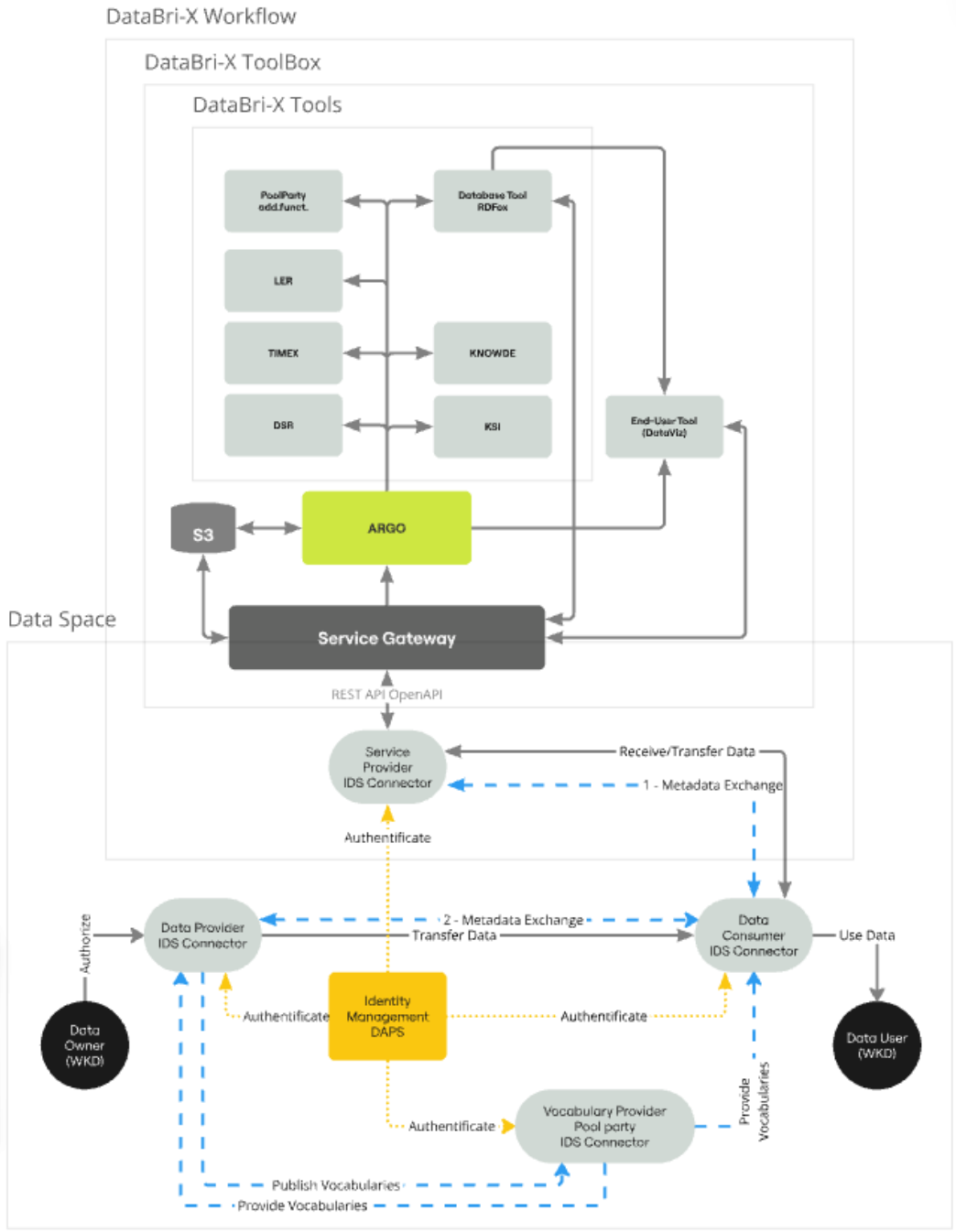


Figure 17 "The Legal Data Space"

4. Conclusion and Outlook

The DataBri-X project represents a significant step towards advancing the European data economy through innovative tools, governance processes, and real-world pilot implementations. By rethinking traditional data lifecycle practices and embracing a holistic, modular approach to data sharing, the project has successfully demonstrated its potential to drive value creation in European Data Spaces.

Key Achievements

Flexible Data Governance Model. The implementation of the JenPlane governance process has provided a dynamic framework for managing complex data lifecycles. By integrating disciplines such as planning, data acquisition, validation, preservation, and integration, JenPlane ensures flexibility, adaptability, and efficiency in data-driven workflows.

Innovative Data Tools. The DataBri-X toolbox has been designed and deployed to facilitate all stages of the data lifecycle, from acquisition and assurance to annotation, analysis, and preservation. These tools address specific challenges such as data interoperability, confidentiality, and metadata enrichment.

Pilot Demonstrations:

- **Telecom Data Space Pilot.** Demonstrated the capabilities of advanced data analytics for telecom providers, enhancing marketing strategies, customer support, and infrastructure management.
- **Energy Data Space Pilot**. Streamlined the design and verification processes of renewable Energy Communities through simulations, facilitating energy grid integration and sustainability.
- **Legal Data Space Pilot**. Created a foundation for a European Legal Data Space, addressing ethical standards, data governance, and the need for unbiased legal data analysis.

The Following Opportunities and Areas of Focus

Scalability Aspects

- **Telecom Data Space Pilot** demonstrated the successful implementation of advanced analytics on real-world datasets. Future advancements will focus on scaling these capabilities to accommodate larger data volumes and more complex use cases, enabling broader adoption across telecom providers.
- **The Energy Community (EC)** simulation workflows demonstrated in the pilot can be scaled to include multiple ECs and more complex energy grid scenarios. Expanding the pilot's reach will allow for more comprehensive simulations, supporting Europe's transition to clean energy.
- **The Legal Data Space** has the potential to scale by integrating more legal jurisdictions and multi-lingual datasets. Future development will focus on expanding its capacity to process larger volumes of legal information while maintaining high precision and usability.

Validation

Further validation of tools and workflows of **Telecom Data Space Pilot** will ensure that sentiment analysis, user clustering, and infrastructure predictions meet industry standards, improving accuracy and reliability.

Energy Data Space Pilot continued validation will focus on hardware-in-the-loop integration and ensuring that simulations align with real-world grid conditions. Accurate verification of EC feasibility will be critical for stakeholder confidence.

Legal Data Space Pilot validation efforts will prioritize refining tools for legal document enrichment, entity recognition, and structural analysis to meet the rigorous demands of legal professionals.

Stakeholders of the Data Space

- **Telecom Data Space Pilot** key stakeholders include telecom providers, SMEs, network infrastructure teams, marketing experts, and customer support departments.
- **Energy Data Space Pilot** key stakeholders include grid operators, renewable energy providers, community energy planners, and technology developers.
- **Legal Data Space Pilot** key stakeholders include law firms, legal departments, public administrations, universities, and technology providers.

Value to the Stakeholders

- **The Telecom Data Space** delivers actionable insights, such as optimized marketing campaigns, improved customer service strategies, and proactive infrastructure management. These insights enhance business efficiency, reduce costs, and improve customer satisfaction.
- **The Energy Data Space Pilot** streamlines the simulation and onboarding processes for ECs, reducing time and cost for grid integration. Stakeholders benefit from improved grid efficiency, increased renewable energy adoption, and enhanced **energy security** for participants.
- **The Legal Data Space Pilot** delivers value through enhanced access to legal knowledge, improved document analysis, and the ability to build smart legal services. By enabling efficient legal research and unbiased decision-making, it fosters productivity and innovation across the legal sector.

Impact and Future Opportunities

The outcomes of DataBri-X contribute to the creation of a trustworthy, FAIR-compliant, and interoperable data ecosystem. The integration of innovative tools and flexible governance strategies empowers stakeholders across sectors - telecommunications, energy, and legal domains - to harness the full potential of their data resources.

Final Outlook

The DataBri-X project lays the foundation for a sustainable, secure, and innovative data ecosystem that aligns with European values. By bridging the gaps in data interoperability, accessibility, and governance, the project paves the way for a thriving data-driven economy where stakeholders can collaborate seamlessly across sectors. Moving forward, the

continued adoption and evolution of DataBri-X methodologies and tools will play a critical role in driving innovation and enabling a more connected, equitable, and responsible future.

References

1. DataBri-X “Data Process & Technological Bricks for expanding digital value creation in European Data 7 Spaces” Project, Grant Agreement No. 101070069, https://databri-x.eu
2. DataBri-X. (31.05.2023). D3.1: Technical Requirements Specification and Development Roadmaps for all Tools. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D3.1_Technical_Requirements_S_v1.4_Submitted.pdf
3. DataBri-X. (31.03.2023). D1.2: Data Management Plan. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D1.2_Data-Management-Plan_SWC_29032023-final.pdf
4. DataBri-X. (28.06.2023). D3.3: Tools & Services, DataBri-X Resource Connector, and DataBri-X IDS Infrastructure and IDS Connector Framework. https://databri-x.eu/wp-content/uploads/2024/07/DataBriX_D3.3.pdf
5. DataBri-X. (26.04.2023). D2.1: Initial Version of Requirement Analysis Result. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D2.1_Initial_Version_of_Requirement_Analysis_Result_v2.0_25042023_NOVA.pdf
6. DataBri-X. (28.09.2023). D5.1: Pilot planning and operational management reports. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D5.1_Pilot-planning-and-operational-management-reports_Submitted.pdf
7. DataBri-X. (29.09.2023). D4.1: Specification and Evaluation Requirements of the DataBri-X toolbox. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D4.1_Specification_and_Evaluation_Requirements_of_the_DataBri-X_toolbox.pdf
8. DataBri-X. (20.07.2023). D3.2: Tools & Services, DataBri-X Resource Connector, and DataBri-X IDS Infrastructure and IDS Connector Framework. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D3.2_Tools_and_Services_DataBri-X_Resource_Connector_and_DataBri-X_IDS_Infrastructure_and_IDS_Connector_Framework_v2.0_20230720_final_Jul2023.pdf
9. DataBri-X. (31.05.2023). D2.2: Initial version of Governance Process Specification, Toolbox Technical Specification, and Pilot Demonstration Specification and repository. https://databri-x.eu/wp-content/uploads/2024/05/DataBriX_D2.2_Initial-version-of-Governance-Process-S_V1.3_Submitted.pdf
10. DataBri-X. (30.08.2024). D5.2: Actual field trials of the Pilot Use Cases, performance evaluation and lessons learned. https://databri-x.eu/wp-content/uploads/2024/10/DataBriX_D5.2_Actual-field-trials-of-the-pilot-use-cases.pdf
11. Georg Rehm, Stelios Piperidis, Dimitris Galanis, Penny Labropoulou, Maria Giagkou, Miltos Deligiannis, Leon Voukoutis, Martin Courtois, Julian Moreno-Schneider, and Katrin Marheinecke. 2024. European Language Grid: One Year after. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6353–6362, Torino, Italia. ELRA and ICCL.

CONTACT

International Data Spaces Association

Emil-Figge-Str. 80
44227 Dortmund | Germany

phone: +49 231 70096 501
mail: info@internationaldataspaces.org

WWW.INTERNATIONALDATASPACES.ORG

international-data-spaces-association