Common European Language Data Space: Development, Current Status, and Future Perspectives

Stelios Piperidis[1], Penny Labropoulou[1], Dimitrios Galanis[1], Khalid Choukri[2], Andrejs Vasiļjevs[3] , Miltos Deligiannis[1], Katerina Gkirtzou[1], Dimitris Gkoumas[1], Athanasia Kolovou[1], Leon Voukoutis[1], Kanella Pouli[1], Maria Giagkou[1], Maria Gavriilidou[1], Katrin Marheinecke[4], Elena Leitner[4], Simon Ostermann[4], Stefania Racioppa[4], Kossay Talmoudi[2], Victoria Arranz[2], Valérie Mapelli[2], Hélène Mazo[2], Fernanda González Campo[2], Shi Yu [2], Aivars Bērziņš[3], Andis Lagzdiņš[3] and Georg Rehm[4]

[1] "Athena" R. C., Institute for Language and Speech Processing (ILSP), Greece
[2]Evaluations and Language Resources Distribution Agency (ELDA), France
[3]Tilde, Latvia
[4]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Corresponding author: spip@athenarc.gr

Abstract

Common European Data Spaces (CEDS) are aimed at creating a single market for data across the EU that will power AI innovation. CEDS cover 14 sectors/domains and will allow secure, trustworthy data/AI models exchange between companies, public administrations and other types of data users/providers. The European Language Data Space (LDS) is part of CEDS and is already made publicly available. The paper presents its technical design and implementation, its governance framework as well as use cases that demonstrate its value. LDS aspires to become part of the future European Language Technology ecosystem.

Keywords: LR infrastructures, language data, data space, language resources

1. Introduction

Recognising the significance of data in fostering global competitiveness and ensuring data sovereignty, the European Union, has outlined a strategic approach to create a unified market for data[1]. The Common European Data Spaces[2] (CEDS) play a pivotal role in this strategy, ensuring that data are made available for use in various economic sectors, while the data owners/producers maintain control over it. Data, considered an essential resource for economic growth, innovation, job creation, and societal progress, underscores the importance of establishing a robust and coordinated data ecosystem.

In the ever-evolving digital landscape, the concept of data spaces has emerged as a crucial framework for enabling secure and trustworthy data transactions among participants, thus offering a boost to the European data economy. According to the DSSC glossary, a data space is defined as an *"Interoperable framework, based on common governance principles, standards, practices and enabling services, that enables trusted data transactions between participants"*(Data Spaces Support Centre, 2026).

As Europe strides towards a more connected and data-driven future, the establishment of the Common European Language Data Space (LDS)[3] represents a significant milestone. By fostering collaboration, innovation, and data sovereignty, the LDS project lays the foundation for a dynamic and thriving European data ecosystem that will propel the continent to the forefront of the global digital landscape.

The paper is structured as follows: Section 2 introduces the architecture and principles that shape the design of LDS. The next two Sections dive deeper into the technical components that support LDS operations. Section 5 explores the dynamics of LDS through the connection of data to processing services. Section 6 is devoted to the Governance Framework, pivotal for the success and sustainability of LDS. Section 7 describes ongoing projects that showcase the potential of LDS for stakeholders. Section 8 presents instruments and activities deployed in support of users and the promotion of LDS. Section 9 describes related work. Finally, Sections 10 and 11 conclude with the current status

[1] https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

[2] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

[3] https://language-data-space.ec.europa.eu/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1179–1188
11–16 May 2026.

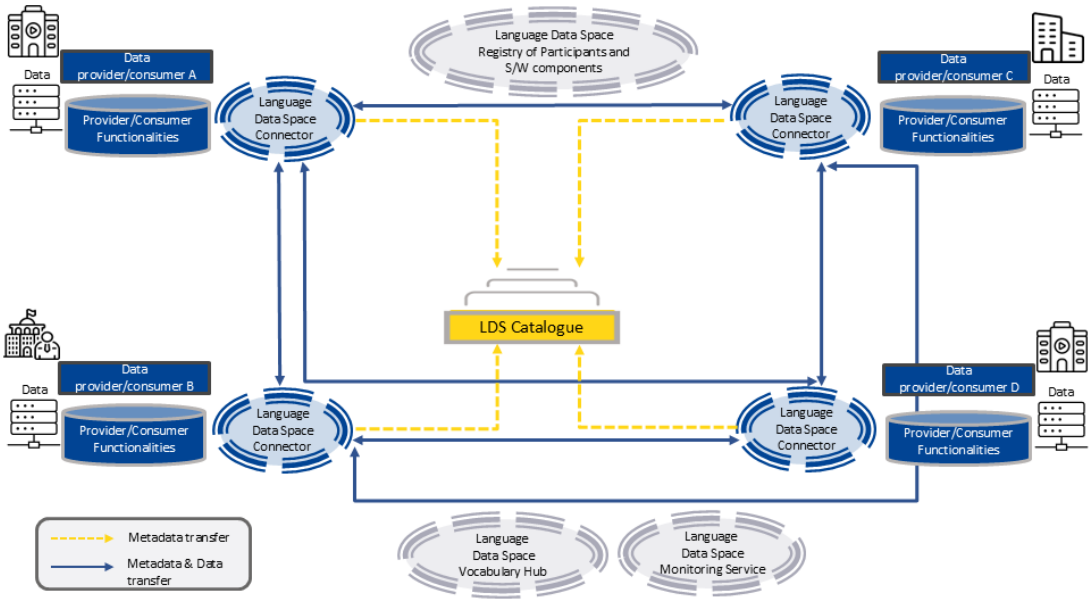

Figure 1: LDS Architecture

and an outlook into the future.

2. Architecture and Design Principles

The three main **principles** in the concept of data spaces are that of (a) **data sovereignty**, i.e., the ability of a person, to exclusively in full power and authority decide on the usage of their own data as an economic asset, (b) **trust**, the possibility to perform data transactions in a secure environment among trusted parties, that are tracked, in full respect of legislation, providing evidence of provenance and transparency, and (c) **interoperability** at the governance, technical and semantic level, maximising (re-)use of data across data spaces.

To serve the above principles, a data space is conceived of and implemented as a set of separate technical components ("participant agents", aka "Connectors") inside which their owners/operators perform all necessary actions related to their own assets and interact with other components (central or other participants' components) through secure communication channels following specified technical protocols when specific criteria are met. The most important protocol is the *Dataspace Protocol (DSP)*[4] that lays the foundations for technical interoperability in data spaces. More specifically, the DSP regulates data sharing transactions in data spaces: metadata descriptions of assets are exchanged between connectors in the form of DCAT[5] catalogues; policies that govern data access and usage are represented in ODRL[6], and electronically negotiated; data are accessed using data transfer APIs/protocols while the interaction between them can be governed using a Data Transfer Control Protocol; finally, logs are generated for each transaction, enabling monitoring and accounting thereof.

The LDS complies with the above principles, adheres to recommended standards, and is built based on, extending and customising relevant technologies to language data requirements in view of developing AI trustworthy systems.

Figure 1 shows the architecture of LDS. The LDS is framed as a decentralised network of organisations that install and operate the **LDS Connector** (Sec. 3), which aggregates functionalities supporting all operations centered around data exchange, from publication of language data to their discovery and actual transfer, as well as the logging of such transactions. Transactions, such as metadata discovery and data flows, are performed Connector-to-Connector. The picture is complemented with four **LDS Central Components**, (Sec. 4) controlled by the LDS Governance Board (GB) (Sec. 6), which aim to facilitate the interactions between participants at different dimensions: the *LDS Registry of Participants*, responsible for the authentication of participants and Connectors, the *LDS Central Catalogue*, giving an overview of the offers provided by the Connectors and acting as an entry point for newcomers, the *LDS Central Monitoring/Logging Service*, tracking transactions between participants under specific conditions, and the *LDS Vocabulary Hub*, contributing to semantic interoperability needs by providing the respective tools and documenta-

[4] https://eclipse-dataspace-protocol-base.github.io/DataspaceProtocol/2025-1/

[5] https://www.w3.org/TR/vocab-dcat-3/

[6] https://www.w3.org/TR/odrl-model/

tion of (meta)data models.

3. LDS Connector

The **LDS Connector** is the main component in LDS, enabling participation in LDS and execution of all automated operations related to data description and exchange, serving both data providers' and data consumers' needs[7]. Upon approval of its application[8] by the GB, an organisation is required to install the LDS connector on its premises, configure and maintain it while participating in LDS.

The LDS connector builds upon, adapts and extends the *Eclipse Dataspace Components (EDC)*[9] *connector*, an open source project, that implements the DSP. The EDC connector provides a variety of basic functionalities for the data workflows, such as the creation, storage and management of metadata descriptions for the data products ("assets"), and licensing terms under which they can be shared ("policies"), their publication in catalogues (as "offers"), the negotiations, policy enforcement, contracts and data transfers. These functionalities have been extended in LDS by (a) developing extensions and adding new endpoints and enhancements, taking advantage of the modular architecture of EDC, and (b) by implementing the additional functionalities in a separate software component (LDS Proxy) set in front of the EDC connector.

3.1. LDS Connector Architecture

The **LDS Connector architecture**, as depicted in Fig. 2, comprises a set of interconnected modules. At the heart of the system (as already mentioned) lies the EDC connector, which uses a database with a predefined generic schema for storing information about assets, policies, contracts, etc., represented in any metadata schema based on RDF(S)/OWL vocabularies. The REST endpoints interchange format (input/output) is JSON-LD. The **LDS proxy** sits in front of EDC and implements additional functionalities, such as (a) advanced search (based on an Elasticsearch index) on the metadata combining free text and faceted querying with filters from the LDS model, (b) metadata validation through the respective SHACL schemas, (c) managing storages, etc.

The LDS connector supports data transfer from/to a **local storage solution**; i.e., it includes an HTTP-based Storage Service that can be used for hosting and serving data when these should be kept locally at the organisation premises. In addition, external storage solutions (Amazon S3 and S3-compatible cloud storages, such as OVHCloud[10]) are currently supported, and others are in the process of being added, by reusing EDC modules. A **local Identity and Access Management (IAM) module** (based on Keycloak) is used for storing all information about individuals affiliated to the organisation that owns/operates it and have access to this connector. Users are identified through the IAM and are authorised to access specific functionalities depending on the user role(s) they have been assigned. When connectors interact with each other, they exchange information only about them and the owner organisations, and not individuals' personal data, thus complying by design with GDPR. A **local logging module**, intended to assist technical and financial administrators to monitor IT performance and business analytics, is employed for tracking the most important events/incidents that take place during the operation of the Connector and extracting statistical reports. As a first step towards the automation of billing procedures and marketplace services, an **invoicing system** (based on SolidInvoice[11]) has been integrated in the Connector. The invoicing module supports the automatic generation of invoices customised to the consumer's and data product's specificities (i.e., price and extra charges such as VAT) and their transfer from provider to consumer's Connector, together with policy enforcement mechanisms that guarantee data transfer only after the provider has confirmed the completion of the payment while all financial transactions take place outside LDS. The Connector comes with a modern **User Interface**, built specifically for LDS. Implemented as a Single Page Application (SPA), its design takes into consideration user-friendliness and performance specifications. Finally, all LDS Components are hidden behind a **reverse proxy** (nginx), for security and load balancing.

3.2. LDS Connector Functionalities

Through the above modules, the LDS Connector offers several functionalities to data providers and consumers. In this subsection, we take a closer look into the way the core workflows of data exchange take place in the LDS context.

Data providers have at their service an editor, guiding non-expert users through the process of describing their assets according to the LDS metadata model (see Section 4.4) in a stepwise approach. The editor provides help tips from the metadata

[7] Although the LDS connector supports both provider and consumer roles, organisations that install them can make use of features intended for one or both roles.

[8] To join the LDS interested organisations can apply at https://language-data-space.eu/registry ui/participant-form.

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

[10] https://www.ovhcloud.com/en/

[11] https://solidinvoice.co/

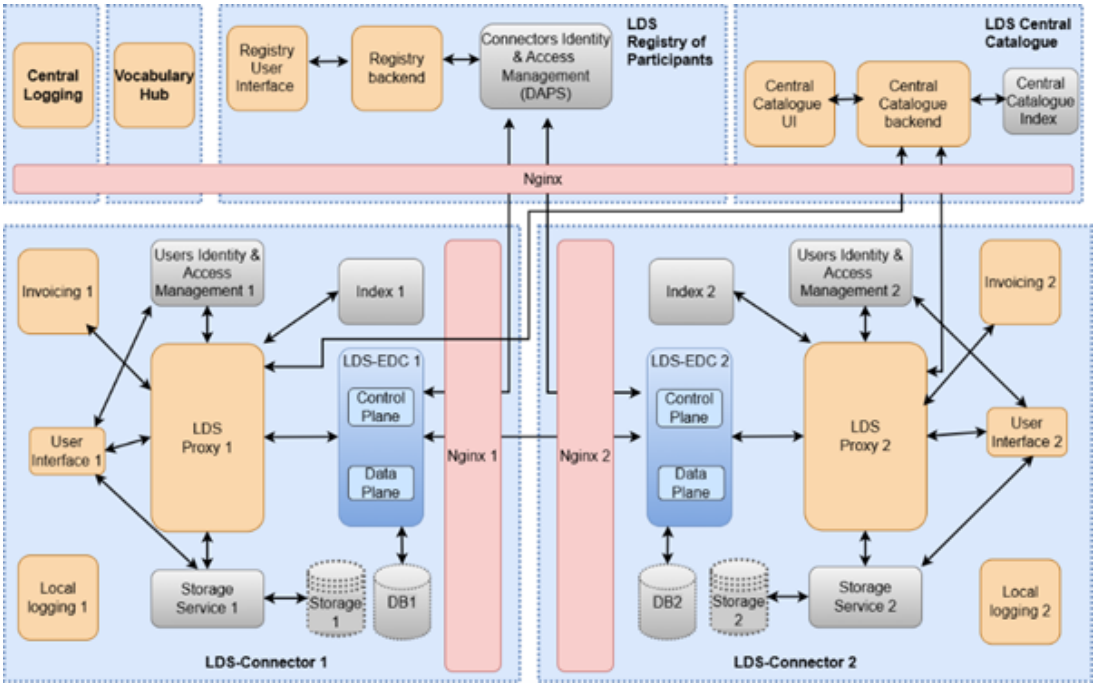

Figure 2: LDS Components

model (e.g., definitions, links to controlled vocabularies, examples). In addition, more expert users can take advantage of an API for the bulk import of metadata descriptions in JSON-LD format. Both the editor and the API are accessible only to users that have editing rights in the specific Connector and integrate a validation module with SHACL rules checking compliance with the LDS metadata model. Users have also the right to manage the metadata descriptions, i.e., update information for the assets and delete them in accordance with the LDS (meta)data lifecycle. Providers can exercise control over their data, by creating and assigning them policies that determine the visibility of their descriptions ("publication policies") and the access and usage of the actual data ("contract policies"). The connector comes prepopulated with a set of policies, namely the "no restrictions" publication policy and several contract policies corresponding to the most popular standard open licences (e.g., Creative Commons family of licences, Apache 2.0). Providers can also add their own custom licences/policies, provided that they are represented in ODRL. This functionality is available in three modes: (a) an API intended for experts, who can upload their policies in JSON-LD format, (b) the LDS standard licence editor and (c) a generic policy editor. The LDS standard licence editor, introduced in V3.0.0, offers providers a ready-to-use legal document created by experts and tailored to the needs of the language data community, as well as its ODRL representation; in addition, the editor guides them to select among a set of pre-defined attributes (e.g., addition of fees, permission/prohibition of derivatives, internal vs. external use, etc.) their preferred options, thus customising the LDS standard licence to their wishes. Finally, the generic policy editor builds upon the notion of "policy class", i.e., an atomic policy template referring to a single specific restriction on data access and/or usage (e.g., geographical or time restrictions, upon payment, etc.) and represented in the form of a ready-to-use ODRL abstract statement (involving *permission*, *prohibition* and/or *duty* in ODRL terminology). A first set of such policy classes, based on an analysis of the licensing terms under which data resources are already provided in existing language data catalogues and the IDSA policy classes (Steinbuss et al., 2021), has been created. The interactive policy editor offers these policy classes to data providers as building blocks that they can easily select, instantiate with their desired values (e.g., adding an amount to the fee for a data asset, or selecting a country where an asset can only be distributed), bundle together and, thus, construct the policies they wish to assign to their data assets. The combination of an asset with a publication and a contract policy creates an offer which is immediately published at the provider's local catalogue and can be crawled by other connectors.

Data consumers can view other connectors' offers (complying with their publication policies) in the form of a catalogue in their own connector. They can browse through the catalogue or search, using filters (e.g., language, keyword, asset type) and/or free text queries. When they find a specific offer

of interest, they can view the contract policy and initiate the so-called "contract negotiation" process, i.e., request to access the data; by doing so, it is assumed that they have accepted the access and usage terms of the offer. This stage involves the communication between the two connectors to perform the negotiation process in full compliance with the DSP specification. This ensures that the actual data transfer can be carried out only when the negotiation has been concluded and computationally enforceable policy terms blocking access to the resource have been enforced (e.g., financial transactions for on-a-fee assets have been completed). Data access and usage policies are evaluated taking into account the context of the transaction itself, the transaction agents as well as the more general setting (e.g., time, for time constraints), leading to a decision as to whether the data transfer can be approved. The policy enforcement rules are implemented within the EDC Connector. Upon completion of a successful negotiation, consumers can proceed to transfer the data into their preferred storage location of their organisation (LDS local storage, cloud infrastructure, etc.), from where the dataset is available to all employees of the organisation, provided they have access to this location[12]. Both provider and consumer Connectors log the respective data transactions as well as the contract between the two parties, recording also the metadata description of the offer and the policy under which it has been acquired. For this feature, extensions have been made to the functionalities offered by the EDC connector.

4. LDS Central Components

4.1. LDS Registry of Participants

The **Registry of Participants** (Fig. 1) hosts a catalogue of the entities participating in the data space, as well as a catalogue of the software components, mainly the Connectors operated/owned by the participants. It contains (a) all information input by the organisation's representative at the application stage that is deemed necessary for identifying the participants and (b) information on the organisations' connectors added at the installation stage. The registry can only be accessed by authenticated and authorised users, namely each organisation's delegated persons (that access only their own organisation's information) and the LDS GB members that are responsible for assessing the application and managing the registry entries in conformance with the LDS Rulebook. The Registry contributes to the trust and security of data spaces: it is responsible for issuing authentication tokens to the participant Connectors thus ensuring that peer-to-peer interactions occur exclusively between authenticated and authorised parties. To support the latter, a central Identity and Access Management (IAM) module (based on Keycloak) has been configured and deployed. For components, a dedicated Keycloak realm (i.e., a logical space for managing a set of users, credentials, roles etc.) has been created. A separate realm in the same IAM is used for the Registry users.

4.2. LDS Central Catalogue

In addition to the catalogues of offers exchanged between participants, the LDS comprises a **Central Catalogue**[13] (Fig. 1) accessible to all, as an additional visibility channel promoting the LDS data offers. The Central Catalogue (CC) includes all offers of a Connector (i.e., without taking into account any access/visibility restrictions that apply to other Connectors), yet in full control of the providers' wishes. Offers of a Connector are by default published at the CC as soon as they are created; still, data providers retain the right to opt-out for all or a subset of their offers, making them available only at local catalogues. The CC does not alter the decentralised nature of LDS; i.e. contract negotiations and data exchanges are allowed only in a peer-to-peer fashion by LDS participants.

The CC is implemented as a web application with free-text and faceted searched capabilities. For its implementation, even though there are open-source components provided by IDSA and the EDC project, we opted to develop it from scratch to meet the desired requirements (e.g., advanced search features) more easily, since major enhancements would be needed for the aforementioned open-source solutions.

A Catalogue of LDS participants is also available at a separate page, with minimal information, i.e., organisation name and country of registration.

4.3. LDS Central Monitoring Service

The LDS **Central Monitoring Service** assists tracking provenance and traceability of offers, by, for instance, keeping information for the transactions made between connectors. The transactions that are currently logged are contract agreements and the respective data transfers (succesful or not). The Central Monitoring System must be a valid LDS participant (for security/privacy reasons), in order to communicate with the organisations' connectors. Access to the logs is allowed only for the members

[12] The current LDS version supports exchange of downloadable data files only; data access through a data service, deployment of models, etc. are under investigation.

[13] https://language-data-space.eu/catalogue/

of the GB (Sec. 6) when this is required, e.g., in the case of dispute.

4.4. LDS Vocabulary Hub

The **Vocabulary Management Module** (a.k.a "vocabulary hub"), enables providers to define, describe, manage and share vocabularies used for the description of their datasets and their structure or contents, thus enhancing data interoperability, one of the crucial desiderata in data spaces. All information in the Vocabulary Hub is shared publicly, under the control of the LDS GB, that specifies and monitors the governance rules and principles for its maintenance and update as well of the contained vocabularies. The LDS Vocabulary Hub is actually composed of a set of separate applications that can be used for viewing and editing the models and vocabularies.

To achieve semantic interoperability, the LDS proposes a common metadata model, which builds upon standard vocabularies, mainly DCAT v.3[14], and domain-specific vocabularies already used in the Language Resources and Technologies community, such as the META-SHARE ontology[15] (Labropoulou et al., 2020; McCrae et al., 2015) for language data and language technologies, and MLDCAT-AP[16] for models. Following common best practices, LanguageDCAT-AP[17] is implemented as an extension/adaptation of DCAT-AP[18], an application profile of DCAT, which is used for the exchange of information about catalogues of datasets and related services. The current stable version (v0.9.2) caters for data resources, and more specifically, corpora and lexical/conceptual resources, as well as language models, and , language technology services, including a minimal set of attributes that help consumers discover offers, while the full set of attributes is in the process of being developed in future versions.

The Language DCAT-AP profile, available in the form of an RDF/OWL ontology, with a set of SHACL rules enforcing cardinality, data types and values of controlled vocabularies, is hosted at a dedicated GitHub repository together with its documentation. The same repository also hosts mappings with other metadata vocabularies (work in progress) and converters. The controlled vocabularies, encoded in SKOS, are also published as autonomous vocabularies[19] visible to all, and available for editing by the Metadata Working Group with the VocBench editing platform[20]. The model and vocabularies, as mentioned in Section 3.2 are used in the editor as tips.

4.5. LDS Central Translation Component

LDS, in full compliance with the principle of language equality, supports multilinguality. Both user interfaces of the Central Catalogue and the LDS Connector, as well as the metadata descriptions of data products are available in all official EU languages. To this end, automatic translation is put into the workflow, while curation of the outputs, where possible and according to a set of governance rules catering for the different types of input, is foreseen.

Upon publication, offers are automatically translated by a central Translation Component which acts as a proxy to the European Commission's eTranslation service[21]. Translation runs asynchronously and all results are stored in the central Component's database, enabling fast subsequent retrieval. Delegated users of the participating organisations may request access to the editing module of the central Translation component in order to review and curate the automatic translations of their own offers.

To ensure improved translation of the technical terminology used in the metadata model, the automatic translation process imports the language versions of the ontology and controlled vocabularies as a separate step. Pre-existing translations for the labels and definitions already included in them, as well as automatically translated ones (through the eTranslation engine again), are imported in the translation component. Collaboration for their curation by terminology experts through the VocBench platform (cf. Section 4.4) and continuous feedback into the translation component is under investigation.

5. Integration with the European Language Grid

The integration of the LDS Connector with the European Language Grid (ELG)[22] (Rehm et al., 2024) demonstrates interoperability between data and service providers: LDS participants can enrich their datasets by using ELG processing capabilities while keeping all transactions within a secure, audit-ready, and GDPR-compliant framework.

[14] https://www.w3.org/TR/vocab-dcat-3/
[15] http://w3id.org/meta-share/meta-share
[16] https://semiceu.github.io/MLDCAT-AP/releases/2.0.0/
[17] https://languagedcat-ap.github.io/LanguageDCAT-AP/
[18] https://semiceu.github.io/DCAT-AP/releases/3.0.0/
[19] https://vocabularies.ilsp.gr/showvoc
[20] https://vocabularies.ilsp.gr/vocbench
[21] https://cor.europa.eu/en/etranslation
[22] https://live.european-language-grid.eu/

The ELG instance of the Connector enables participants to process LDS datasets using ELG services such as translation, speech recognition, or text analytics directly through the LDS Connector. The ELG Connector maps ELG API endpoints to LDS and manages secure data exchange via the LDS policy enforcement and identity management components. ELG metadata (service descriptions, parameters, Input/Output schemas) are mapped into LDS metadata. Service calls are logged to the LDS monitoring module to ensure traceability and compliance.

The integration of the LDS Connector with the ELG is achieved through a component called Service Agent, which is designed in a generic and extensible manner. This design enables not only the ELG, but also other service providers, to interoperate with the LDS through a set of common APIs.

6. Governance Framework

The establishment of an effective **governance framework** is a cornerstone for the success and sustainability of LDS. As a complex, multi-stakeholder ecosystem involving data providers, consumers, technology developers, and public institutions, LDS requires clear, transparent, and enforceable rules to ensure trust, fairness, and accountability. Governance is essential not only to guarantee compliance with European legislation and ethical standards but also to balance the interests of diverse actors, prevent market distortion, and uphold the principles of data sovereignty and interoperability that underpin the European Data Strategy.

The LDS governance framework adopts a multi-layered model distinguishing between institutional and operational bodies, respectively the LDS Governance Board (GB) and the LDS Management Board (MB). During the procurement phase, the European Commission (EC) acts as the GB, defining and supervising the overall strategy, approving participation rules, and validating the General Terms and Conditions. In addition, the EC is entrusted with approving the applications of prospective participants, ensuring that all entities admitted to LDS meet the eligibility, legal, and ethical requirements set out in the governance framework. The LDS consortium assumes the role of MB, responsible for the day-to-day management and enforcement of the EC's decisions. This includes the auditing and onboarding of participants, the operation of the Central Components (Sec. 4), the compilation and maintenance of the **LDS Rulebook**, and the continuous alignment of operational procedures with European legal and ethical frameworks such as the *General Data Protection Regulation (GDPR)*, the *Data Act*, the *Data Governance Act (DGA)*, and the *AI Act*.

Complementing the institutional framework, the LDS Rulebook and the General Terms and Conditions (GT&C) form the operational backbone of the governance scheme. The LDS Rulebook consolidates all applicable rules, policies, and procedures governing the participation and conduct of LDS stakeholders, while the GT&C formalise the rights, obligations, and liabilities of participants. These documents are living instruments, regularly updated to reflect new legislative developments and technical evolutions. They address aspects such as eligibility criteria, licensing and intellectual property management, data protection and privacy, dispute resolution, and termination procedures. Beyond regulatory oversight, the governance framework also plays an essential role in facilitating data sharing within LDS. By establishing clear rules for the participants' interactions, and the way they conduct LDS operations, observing adherence to such rules, offering clear appeal and redress procedures, and accountability mechanisms, the governance framework not only safeguards compliance and trust but also actively promotes the circulation and reuse of language data, both within LDS and across sectors and borders, including with other data spaces, thereby strengthening the collaborative ecosystem of the European data economy.

7. Use Cases

Several Proof-of-Deployment Concept Projects (PoDCPs) are being implemented to serve as example use cases demonstrating the value of participation in LDS.

The PoDCPs validate LDS functionality, technologies, governance principles, and data sharing mechanisms across public and private sectors. These pilots connect national and sectoral stakeholders to the LDS infrastructure via the LDS Connector, testing interoperability, data sovereignty, and value creation in real-world settings.

In cooperation with the Culture Information Systems Centre of Latvia (CISC), the operator of Latvia's national language-technology platform HUGO.LV (Skadiņš et al., 2020), the project integrates CISC into the LDS to enable secure data exchange and the fine-tuning of a Large Language Model (LLM) using national datasets and resources shared by other LDS participants. The resulting TildeOpen LLM (Bergmanis et al., 2026) fine-tuned for machine translation (MT) task shows significant quality improvements compared to bidirectional neural MT systems currently used in HUGO.LV.

The National Library of Latvia (NLL) PoDCP explores the use of customised LLMs for bibliographic metadata generation and multilingual

search. This will support cataloguers in processing Latvian-language periodicals by automating structured metadata extraction. The project also assesses legal, organisational, and technical aspects of library data participation, providing a replicable model for other European libraries.

Another example is the German online community platform Gutefrage that has collected a large corpus of user-generated question–answer discussions covering a wide range of everyday topics. Gutefrage, like many other online platforms, is looking for alternative business and revenue models beyond SEO traffic. They are interested in the monetisation potential of their data in the context of LDS, as it can be very useful for AI developers to train and evaluate question answering systems, conversational assistants, and domain-adapted LLMs. In this use case, selected anonymised discussion data are shared through the LDS infrastructure and can be used by AI developers. The pilot demonstrates how community platforms can use LDS infrastructure to contribute valuable conversational language data while maintaining control over data access and usage conditions. It also illustrates the potential of LDS to support the development of AI systems grounded in real-world user interactions.

8. Promotion, Dissemination, Collaborations and Coordination

8.1. Support for Users

To ensure efficient participant support and to reinforce the principles of transparency and trust, LDS operates three dedicated helpdesks: **legal, technical**[23], and **business development**[24], which together form an integral part of its governance and service infrastructure.

The **Legal Helpdesk** provides non-binding guidance on questions related to data collection, licensing, intellectual property rights, and data protection compliance under EU regulations. It advises participants on licensing policies and the use of the LDS policy editor, assisting them in selecting or customising appropriate licence terms for their offers. The **Technical Helpdesk** offers first-line support for onboarding, connector configuration, and general troubleshooting. It provides guidance throughout the process of publishing and accessing language data within LDS. The **Business Development Helpdesk** assists organisations in exploring data monetisation opportunities, offering advice on pricing strategies and models, and market positioning for language data and services. It also promotes the uptake of LDS services across sectors and encourages sustainable business practices aligned with the European data economy.

For the support of users, LDS offers a **documentation toolkit**, consisting of (a) user manuals continuously updated to be in sync with the LDS versions[25], (b) demo videos and short video tutorials[26], (c) a dedicated FAQs section[27] on legal, technical and general issues at the LDS website and (d) a Swagger documentation for technical experts that wish to use LDS Connector APIs.

8.2. Promotion and Dissemination

The LDS **website**[28], maintained on the Europa Web Publishing Platform, is the main communication hub for LDS, while communication and promotion activities strengthen the visibility and outreach of the LDS. The consortium has developed its coordinated communication strategy across **social media**, the project website, newsletters, and event-related channels. Targeted campaigns focusing on specific themes, “awareness”, “testimonial” and “How to”, with the added use of short videos, have a significant impact.

The LDS consortium initiated an **LDS Interest Group (IG)** comprising of individuals and organisations[29] who would like to get more engaged in LDS developments and influence the LDS progress. The IG receives regular mailings containing more technical and more detailed information (the so-called "digests"), and are invited to dedicated technical or stakeholder workshops, as well as to 1-2 IG Meetings where the latest state of play of the LDS infrastructure as well as other new developments within LDS are reported. The participation in the LDS IG is open to everyone and requires only a sign-up in the online form.

LDS collaborates with all relevant projects and initiatives in the wider data space landscape (many of which are mentioned in this paper), especially DSSC and SIMPL, but also LLM-developing projects including, among others, HPLT[30].

[23] https://language-data-space.ec.europa.eu/help/lds-legal-and-technical-helpdesks_en

[24] https://language-data-space.ec.europa.eu/help/lds-business-development-helpdesk_en

[25] https://docs.language-data-space.eu/

[26] https://www.youtube.com/@LangDataSpace

[27] https://language-data-space.ec.europa.eu/help/faq_en

[28] https://language-data-space.ec.europa.eu/

[29] As of mid-March 2026, the Interest Group has 296 indivudual members out of 220 organisations.

[30] https://hplt-project.org/

9. Related Work

The LDS follows a long line of language data and language technology sharing infrastructures and catalogues (Piperidis et al., 2023) with different designs and architectures, targeting different audiences, and/or with different focus on the contents. Most notable among them, META-NET[31] with META-SHARE[32] (Piperidis, 2012), CLARIN[33] with its various national repositories (Fišer and Witt, 2022), ELRC (Lösch et al., 2018) with ELRC-SHARE (Piperidis et al., 2018), the ELRA catalogue of resources[34], and the most recent European Language Grid[35] (Rehm, 2023; Rehm et al., 2024) and European Language Equality[36] (Rehm and Way, 2023). The design and implementation of the LDS is informed by the above activities, especially since the consortium partners have been involved in most of them. At the same time, platforms for sharing data and models built for the neighbouring Machine Learning and AI communities, such as Hugging Face[37], Kaggle[38], AI-on-demand[39], have also influenced the LDS design.

As aforesaid, the LDS is one of the CEDS funded by the EC, covering 14 sectors, and thus builds upon data space principles and technologies. Related initiatives are the IDSA[40], leading the technical advances, Gaia-X[41], with a focus on the trust framework, the Big Data Value Association[42], and, most importantly, the EU funded projects aiming to support and facilitate the interoperability across the CEDS: (a) DSSC[43] conducts surveys, collects catalogues of standards and tools, promotes collaboration and develops guidelines, the most important of which is the blueprint for data spaces; (b) Simpl[44] intends to develop an open source, smart and secure middleware platform that supports data access and interoperability among European data spaces

The LDS technical design has been inspired from the Mobility[45] and Catena-X Automotive[46] Dataspaces that do not belong to the CEDS and are funded by the German government. Both of them have open-sourced their code, at least partially.

LDS goes beyond previous data and language technology services cataloguing and sharing platforms, by focusing on data sovereignty and interoperability. Secure exchange of datasets is performed in a by-design decentralised, peer-to-peer mode, adding governance and compliance to legal frameworks, and opening new channels for the repurposing and monetisation of data from other domains and communities as well. LDS is designed as an ethical and legally-compliant marketplace ecosystem for language data and AI models across multiple sectors.

10. Current Status

At the time of writing, the LDS comprises 73 participating organisations, i.e., organisations that have gone through the application process and been formally approved by the EC. Individuals affiliated to them are testing the current version of the LDS infrastructure, either through a demo environment or by installing their own connector.

The **LDS infrastructure v3.0.0**, launched in January 2026, is the first stable one, following two beta versions. Extending the capabilities of previous versions and taking into account the feedback received from its testers and the LDS Interest Group, this version supports all functionalities described in section 3.2, albeit with some restrictions (e.g., limited deployment of services, continuous enrichment of policy classes, etc.). All central components have been implemented, while the LDS Connector includes all modules described in Sec. 3.

11. Summary and Future Work

The LDS project has already made available the European Language Data Space infrastructure. The LDS Connector and Central Components have been used and tested by a significant number of organisations. In addition, the project has framed a solid and coherent Governance framework that guarantees data sovereignty, compliance with EU legislation as well as secure and trustworthy operation. The project has guaranteed a one year extension of its current contract for supporting further enhancements and continues the development of the infrastructure in view of the next releases (July 2026, January 2027). Furthermore, all possible options are investigated for safeguarding its long-term sustainability, e.g., via ALT-EDIC. Through continuous dissemination events and collaborations, LDS attracts new participants and its catalogue of language data, models (and, soon to be added, language technologies) is continuously growing.

[31] http://www.meta-net.eu
[32] http://www.meta-share.org/
[33] https://www.clarin.eu/
[34] https://catalogue.elra.info/en-us/
[35] https://live.european-language-grid.eu/
[36] https://european-language-equality.eu
[37] https://huggingface.co/
[38] https://www.kaggle.com/
[39] http://aiod.eu/
[40] https://internationaldataspaces.org/
[41] https://gaia-x.eu/
[42] https://bdva.eu/
[43] https://dssc.eu/
[44] https://simpl-programme.ec.europa.eu/
[45] https://mobility-dataspace.eu/
[46] https://catena-x.net/

12. Acknowledgements

The Common European Language Data Space is funded by the European Union through the contract LC-01936389.

13. Bibliographical References

Toms Bergmanis, Martins Kronis, Ingus Jānis Pretkalniņš, Dāvis Nicmanis, Jeļizaveta Jeļinska, Roberts Rozis, Rinalds Vīksna, and Mārcis Pinnis. 2026. TildeOpen LLM: Leveraging Curriculum Learning to Achieve Equitable Language Representation. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC 2026)*, Spain. European Language Resources Association (ELRA).

Data Spaces Support Centre. 2026. Dssc blueprint.

Darja Fišer and Andreas Witt, editors. 2022. *CLARIN: The Infrastructure for Language Resources*. De Gruyter, Berlin, Boston.

Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitrios Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva. 2020. Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3421–3430, Marseille, France. European Language Resources Association (ELRA).

Andrea Lösch, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith. 2018. European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

John P. McCrae, Penny Labropoulou, Jorge Gracia, Marta Villegas, Víctor Rodríguez-Doncel, and Philipp Cimiano. 2015. One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. In Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, editors, *The Semantic Web: ESWC 2015 Satellite Events*, Lecture Notes in Computer Science, pages 271–282. Springer International Publishing.

Stelios Piperidis. 2012. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou. 2018. Managing Public Sector Data for Multilingual Applications Development. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stelios Piperidis, Penny Labropoulou, Dimitrios Galanis, Miltos Deligiannis, and Georg Rehm. 2023. *The European Language Grid Platform: Basic Concepts*, pages 13–36. Springer International Publishing, Cham.

Georg Rehm, editor. 2023. *European Language Grid – A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer International Publishing, Cham.

Georg Rehm, Stelios Piperidis, Dimitrios 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 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy. ELRA.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer.

Raivis Skadiņš, Mārcis Pinnis, Artūrs Vasiļevskis, Andrejs Vasiļjevs, Valters Šics, Roberts Rozis, and Andis Lagzdiņš. 2020. Language Technology Platform for Public Administration. In *Human Language Technologies–The Baltic Perspective*, pages 182–190. IOS Press.

Sebastian Steinbuss, Andreas Eitel, Christian Jung, Robin Brandstädter, Arghavan Hosseinzadeh, Sebastian Bader, Christian Kühnle, Pascal Birnstill, Gerd Brost, Gall, Fabian Bruckner, Norbert Weißenberg, and Benjamin Korth. 2021. Usage Control in the International Data Spaces.