Common European Language Data Space

Georg Rehm[1], Stelios Piperidis[2], Khalid Choukri[3], Andrejs Vasiļjevs[4], Katrin Marheinecke[1], Victoria Arranz[3], Aivars Bērziņš[4], Miltos Deligiannis[2], Dimitrios Galanis[2], Maria Gavriilidou[2], Maria Giagkou[2], Katerina Gkirtzou[2], Dimitris Gkoumas[2], Annika Grützner-Zahn[1], Athanasia Kolovou[2], Penny Labropoulou[2], Andis Lagzdiņš[4], Elena Leitner[1], Valérie Mapelli[3], Hélène Mazo[3], Simon Ostermann[1], Stefania Racioppa[1], Mickaël Rigault[3] and Leon Voukoutis[2]

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

Abstract

The Common European Language Data Space (LDS) is an integral part of the EU data strategy, which aims at developing a single market for data. Its decentralised technical infrastructure and governance scheme are currently being developed by the LDS project, which also has dedicated tasks for proof-of-concept prototypes, handling legal aspects, raising awareness and promoting the LDS through events and social media channels. The LDS is part of a broader vision for establishing all necessary components to develop European large language models.

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

1. Introduction and Objective

The European Union has been implementing its Data Strategy since 2019.[1] A key component of the industrially-oriented single market for data, in which "data can flow within the EU and across sectors, for the benefit of all", "European rules […] are fully respected" and in which "the rules for access and use of data are fair, practical and clear", is the vision of establishing "interoperable data spaces" for "pooling European data in key sectors". The EC (2022) describes an initial, rather coarse-grained concept how these data spaces are supposed to be set up and operated including the relevant legislation (also see Nagel and Lycklama, 2021). The document also lists a number of 'official' EU data spaces targeting sectors such as manufacturing, mobility, health, finance, energy, agriculture and skills. The Common European Language Data Space (LDS), funded through a procurement contract in the Digital Europe Programme (DEP),[2] is one of these official EU data spaces.[3]

The development of a data space for language data has recently become increasingly relevant due to the importance of large language models (LLMs) for industrial and academic applications and their dependence on vast amounts of language data for training models, especially when considering the increasing push of the EU towards digital sovereignty. LDS is one crucial part of a bigger initiative towards more European independence and increased European participation in the global LLM landscape (including research, development, innovation, application, deployment and also monetisation).

The main objective of the LDS initiative is the development of the Language Data Space while at the same time supporting the development of the Alliance for Language Technologies EDIC (ALT-EDIC, see Section 3 for more details).

The LDS consortium consists of private/public and commercial organisations with a track record in language technologies research and development (including technology development and service provision, data collection, platform and infrastructure development and processing aspects). Among the subcontractors are various organisations throughout Europe who are tasked with organising country workshops to promote the initiative to local and regional stakeholders. LDS is further supported by legal experts and an agency for additional help for communication and dissemination activities (3pc GmbH, Germany). The Big Data Value Association (BDVA) helps with outreach activities while CLARIN ERIC establishes a bridge between LDS and the CLARIN network.

[1]https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
[2]https://digital-strategy.ec.europa.eu/en/activities/digital-programme
[3]https://language-data-space.ec.europa.eu

LREC-COLING 2024, pages 3579–3586

This paper is structured as follows: Section 2 summarises the methodology, concentrating on governance, the technical infrastructure, communication and promotion as well as coordination and support. Section 3 describes related work. Finally, Section 4 provides a summary and outlook.

2. Methodology and Main Tasks

To illustrate the main work areas and priorities of the project, the LDS initiative is structured into 14 tasks: T1 Coordination and Support; T2 Centre of Excellence for Language Technologies (CELT); T3 Multi-Stakeholder Governance Body (CELT+, i. e., the LDS User Group); T4 Multi-Stakeholder Data and Services Governance Scheme; T5 Implementation of the Multi-Stakeholder Data and Services Governance Scheme; T6 Sustainable Language Data Ecosystem Blueprint; T7 Implementation of the Sustainable Language Data Ecosystem Blueprint; T8 Language Data Space Deployment; T9 Proof-of-Deployment-Concept Projects; T10 Event Organisation and Management; T11 Promotional Activities through Conference Attendance; T12 Promotional Activities through Social Media and other Information Channels; T13 Language Data Space Website; T14 Data Protection Compliance. The contract runs for a period of 36 months after which the project can be renewed by the EU for an additional 12 months. The main tasks are grouped together and described in the following subsections.

2.1. Governance Bodies

LDS is currently setting up two boards, CELT (Centre of Excellence for Language Technology[4]) and the LDS User Group, formerly known as CELT+. The CELT will eventually consist of representatives of Member States and DEP-associated countries (i. e., Iceland, Liechtenstein, Norway, Ukraine). There is an intentional overlap between the CELT and the representatives of the EU Member States active in the ALT-EDIC (see Section 3), the development of which is supported by the LDS contract.

LDS is a multi-stakeholder initiative and, hence, the LDS User Group, established in March 2024 with an initial core seed of 39 member organisations, consists of different stakeholder types: members representing European companies, European associations, public administrations and NGOs as well as other organisations. Over time, the LDS User Group will grow, especially with regard to a larger number of industry representatives covering various sectors. As many industry sectors as possible and also all EU Member States including DEP-associated countries will be represented to assure that the LDS takes into account as many perspectives as possible. For instance, the LDS User Group will comprise publishing houses, media, automotive industry, e-commerce, etc. Like with the CELT, there is an intentional overlap between the LDS User Group and the industrial consortium that will support the ALT-EDIC in the implementation of its plans. Within LDS, the CELT ensures that all national requirements, visions and strategies are properly reflected in the overall LDS concept and roadmap, as developed by the LDS consortium, while the LDS User Group helps with the implementation and validation of the strategic concepts.

2.2. Governance Framework

The project develops a fair and efficient scheme for the governance of the multi-stakeholder data and services to be offered through the soft LDS infrastructure, coordinated by the CELT in collaboration with the LDS User Group (see Section 2.1). Here, governance is part of a more global set of pillars that would allow EU stakeholders to access resources for various purposes.

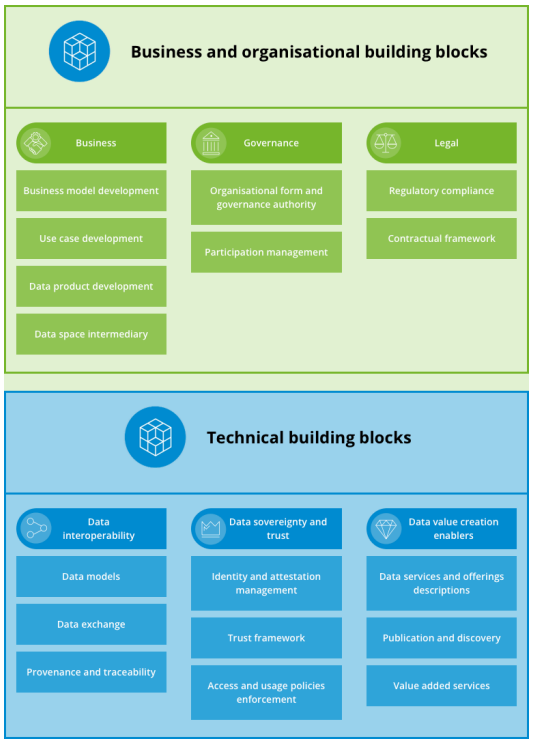


Figure 1: DSSC Data Spaces Building Blocks

In order to harmonise all data spaces supported by the EC under the DEP, the project Data Spaces Support Centre (DSSC)[5] develops a blueprint taking into account feedback and input from the var-

[4]This name is likely to change.

[5]https://dssc.eu

ious data space initiatives. These building blocks are considered by LDS for its ongoing development (Figure 1, see Guilloud et al., 2024) covering organisational and business building blocks as well as technical building blocks.

The governance scheme will address technical and non-technical facets (legal, operational, technical etc.). The consortium will also adhere to all regulations that may impact the governance scheme, in particular those that are related to the operation of collecting and sharing language data and services. The governance scheme will also cover the infrastructure, cataloguing and sharing of language data and services. LDS will ensure that all requirements expressed by the CELT, LDS User Group, and the EC are taken into account regarding the governance structure. The scheme will regulate the trusted onboarding of data space participants (data providers, data owners, data consumers). This process will follow several principles, in particular compliance with EC requirements as well as EU regulations (GDPR, AI Act, Data Act, etc.). Crucially, the legal framework will ensure trust between all data space participants (Guilloud et al., 2024).

The governance scheme will require each LDS participant to use minimal metadata descriptions to ensure discoverability and facilitate interoperability of their assets (see Section 2.3). The metadata records of each participant's LDS connector (see Section 2.3) can be harvested to maintain a central catalogue through which other LDS participants can discover concrete offers, i. e., one or more data sets offered by another participant, including licensing conditions, terms of use, price etc. With regard to all governance topics, the governance scheme will benefit from the partners' expertise in running infrastructures (ELRA[6], ELG[7], ELRC-SHARE[8], META-SHARE[9] etc.).

The implementation of the governance policies will include adherence to data protection protocols, automation of workflows, and compliance with privacy regulations. Key tasks involve user and identity management (applicable to both human and machine users) as well as data access policies, perhaps eventually extending to data validation, storage, benchmarking, curation, and exchange. The architecture will ensure a decentralised and scalable integration of all deployed LDS connectors while adhering to governance, security, and interoperability standards across the LDS. Security and trust will be ensured by encrypted data exchange protocols and interoperable identity solutions like OpenID Connect[10] and OAuth 2.0,[11] perhaps including eIDAS[12] as an identity provider. The technical framework will be open-source and based on standards, ensuring compatibility and enabling widespread adoption. Standards such as Open Digital Rights Language (ODRL)[13] and Data Privacy Vocabulary (DPV)[14] will improve adaptability and compliance. ODRL is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. DPV provides a vocabulary for the expression of machine-readable metadata about the use and processing of personal data based on legislative requirements such as GDPR. This comprehensive approach embodies a vision for an efficient, secure, and interoperable LDS, strategically positioned to elevate the European Data Space ecosystem.

2.3. Technical Infrastructure

The main principle in the concept of data spaces is that of data sovereignty, i. e., the ability of a person, natural or legal, to exclusively decide, in a sovereign way, on the usage of their own data as an economic asset (Section 2.2). As a result of this principle, the LDS gives owners/providers full control regarding access to their data, also enabling data transactions including monetary transactions. The LDS will support tracking of data provenance and lineage and it will enforce access and usage policies formally codified as contracts established in a trustful environment between participants. The typical operation technically supported in LDS is that of exchange of information about assets (this term includes data, data products and services as well as related offerings) and exchange or transfer of such assets between trusted participants. An asset in LDS can be any type of language-related dataset (corpus, lexical/conceptual resource, language model) or a language processing service.

Participants will be technically represented through software artifacts called connectors; the LDS connector is developed based on the Eclipse Dataspace Components (EDC).[15] A connector is the indispensable component for all participants, data providers and data consumers, that enables the documentation of language data, the cataloguing of assets on offer by the connector's owner, the contract-based agreement for the exchange

[6] https://www.elra.info
[7] https://www.european-language-grid.eu
[8] https://www.elrc-share.eu
[9] http://www.meta-share.org
[10] https://openid.net/connect/
[11] https://oauth.net/2/
[12] https://www.eid.as
[13] https://www.w3.org/TR/odrl-model/
[14] https://w3c.github.io/dpv/dpv/
[15] https://projects.eclipse.org/projects/technology.edc

of data, the actual transfer or processing of data and the logging of all transactions. Thus, the LDS can be conceptualised as the aggregation of all its connectors, connected in a peer-to-peer fashion. Each connector is equipped with a local catalogue in which this provider's offerings are published and available for querying or crawling by other connectors (Figure 2). In addition, a central catalogue will give an overview of all offerings publicly available. The LDS will eventually be a marketplace allowing users, providers and consumers to perform, manage and monitor their commercial transactions and exchange data in full respect of the contractual agreements made between the parties.

To ensure accurate data usage by individuals, components and services, interoperability is key. At the technical level, common data formats and standard interfaces/APIs for the exchange/sharing of data deemed appropriate for the Language Technology domain will be recommended based on current best practices and protocols broadly used, upon consultation with the CELT and LDS User Group. At the semantic level, common vocabularies, including ontologies, data models, schema mappings and converters will be used to describe and annotate data and services, facilitating understanding of the data. LDS builds upon standard vocabularies, such as DCAT,[16] as well as popular domain-specific vocabularies such as the META-SHARE ontology and also on approaches developed for ELG (Labropoulou et al., 2020; Rehm et al., 2021).[17] Mappers and converters to other metadata schemas will be developed to bridge with other infrastructures (e. g., ELG, ELRC, CLARIN). In addition, a special component, the vocabulary hub, will allow providers to define, describe, manage and share other vocabularies used for the description of their data. The implementation of the infrastructure will be compliant with the Data Space Protocol (IDSA, 2024) and the DSSC recommendations, while adapting and extending them to the practices and requirements of the LDS target communities.

2.4. Communication and Promotion

The LDS initiative is meant to develop into the leading European marketplace for language data, which is why communication and promotion play an important role. The communication and dissemination plan foresees a mix of instruments and channels. It pursues the following objectives: 1. Promoting the LDS initiative to stakeholder groups to raise awareness and collect feedback for shaping the LDS infrastructure; 2. Encouraging the use and uptake of the LDS amongst European actors from the private and public sectors; 3. Promoting all relevant project activities through workshops and conferences;[18] 4. Providing a platform for community building and exchange. A key goal is to build, expand, interconnect and firmly establish a community around LDS and its content and functionalities. To reach a significant number of stakeholders, we benefit from related networks (e. g., ELG, ELE, ELRC, DSSC) and our subcontractors (e. g., BDVA) who will help spread the word and also help amplify the LDS newsletter, social media channels and website.

2.5. Coordination and Support

Three tasks support LDS's core activities. Proof-of-deployment concept projects will validate the value chain enabled by the LDS and legal support is provided regarding aspects such as data protection compliance. The project management task also includes the technical and legal helpdesks, which provide support to the users of the LDS.

3. Related Work

Throughout the decades, a number of Natural Language Processing (NLP) and Language Technology platforms has been developed (Piperidis et al., 2023). The partners of the LDS consortium have been involved in several of these, including European Language Grid (ELG, Rehm, 2023; Rehm et al., 2024) and European Language Equality (ELE, Rehm and Way, 2023),[19] ELRC[20] (Lösch et al., 2018) with ELRC-SHARE (Piperidis et al., 2018), META-NET[21] with META-SHARE (Piperidis, 2012) and CLARIN[22] with its various national repositories (Fišer and Witt, 2022). These activities inform the design of the LDS infrastructure, especially ELG and its recent updates, which aim at the integration into Gaia-X[23] as well as into the International Data Spaces Association (IDSA)[24] data space paradigm. In the EU project DSSC, Gaia-X, IDSA, the Big Data Value Association (BDVA)[25] and FIWARE[26] (i. e., the Data Spaces Business Alliance, DSBA, 2023)[27] and

[16] https://www.w3.org/TR/vocab-dcat-3/
[17] http://w3id.org/meta-share/meta-share

[18] See, for example, the LDS Technology Workshop "Legislation and regulations for data spaces: an environment for the development of a European Data Market" (29 January 2024), https://language-data-space.ec.europa.eu/events/.
[19] https://european-language-equality.eu
[20] https://lr-coordination.eu
[21] http://www.meta-net.eu
[22] https://www.clarin.eu
[23] https://gaia-x.eu
[24] https://internationaldataspaces.org
[25] https://www.bdva.eu
[26] https://www.fiware.org
[27] https://data-spaces-business-alliance.eu

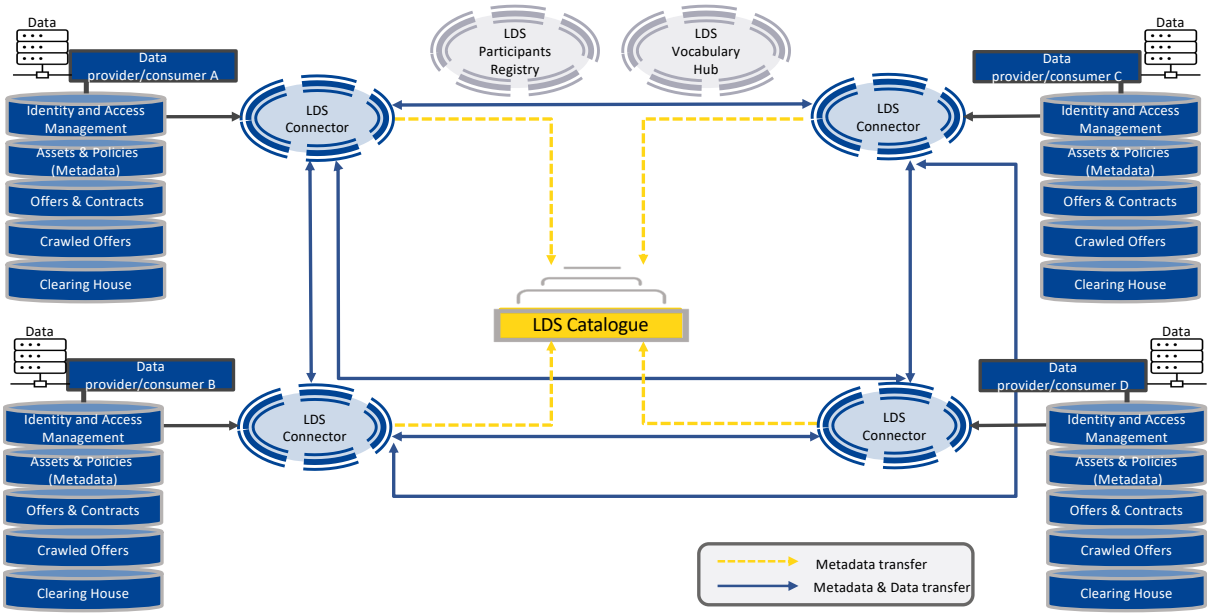


Figure 2: LDS technical infrastructure

other European organisations develop, together with all other data space initiatives (including LDS), a common blueprint for data spaces, version 1.0 of which was published in March 2024 (Guilloud et al., 2024). This blueprint is meant to be used as a joint approach for all common European data spaces. The EU framework contract Simpl started in early 2024.[28] It will develop a joint middleware for all common European data spaces. As technical, legal and operational interoperability are mission-critical, LDS collaborates with all relevant data space initiatives, especially the Trusted European Media Data Space (TEMS).[29] Additionally, we expect *national* language data spaces to emerge in the future. LDS will attempt to collaborate with these to achieve full interoperability amongst all emerging data spaces.

In a parallel development, the EU established the European Digital Infrastructure Consortium (EDIC) as a new legal entity type.[30] EDICs enable EU Member States to set up multi-country projects with regard to the Digital Decade Policy Programme 2030. One of the first is the Alliance for Language Technologies (ALT-EDIC), which seeks to establish a Europe-wide technology centre to develop a common infrastructure for NLP and LLMs.[31] While the EDIC is governed and financed by the participating EU Member States, industry, academia, NGOs and other stakeholders will also play a role. LDS collaborates with the ALT-EDIC and the EC to synchronise both activities.

4. Summary and Future Work

The LDS project is developing the Common European Language Data Space, and it has recently concluded its first year. The LDS User Group has started its work in March 2024, and important decisions with regard to setting up the CELT as its strategic steering committee have been made. The first version of the technical blueprint of the infrastructure has been completed in early 2024 and all communication channels have been set up. We expect the first prototype installations of the LDS connector to be deployed in the summer of 2024 (see Figures 3 and 4 for impressions of the current version). We are very confident that the LDS – in close collaboration with the ALT-EDIC – will have a very strong positive impact on the overall language data situation in Europe including the availability of data for under-resourced languages and making available language data from various industries and sectors.

5. Acknowledgements

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

[28] https://digital-strategy.ec.europa.eu/en/policies/simpl

[29] https://tems-dataspace.eu

[30] https://digital-strategy.ec.europa.eu/en/policies/edic

[31] https://language-data-space.ec.europa.eu/related-initiatives/alt-edic

6. Bibliographical References

DSBA. 2023. Technical Convergence Discussion Document. Data Spaces Business Alliance, April 2023.

EC. 2022. Commission staff working document on common european data spaces.

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

Geraud Guilloud, Mirjam Huisintveld, Tonia Sapia, and Clara Pezuela, editors. 2024. *Data Spaces Blueprint 1.0, March 2024*. Data Spaces Support Centre (DSSC).

IDSA. 2024. *Data Space Protocol 2024-1*. International Data Space Association.

Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris 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 *Proc. of the 12th Language Resources and Evaluation Conf. (LREC 2020)*, pages 3421–3430, Marseille, France. 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 *Proc. of the 11th Int. Conf. on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

Lars Nagel and Douwe Lycklama. 2021. Design Principles for Data Spaces. Position Paper. Technical report, Berlin.

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey. 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. ELRA.

Stelios Piperidis, Penny Labropoulou, Dimitris Galanis, Miltos Deligiannis, and Georg Rehm. 2023. The European Language Grid Platform: Basic Concepts. In Georg Rehm, editor, *European Language Grid: A Language Technology Platform for Multilingual Europe*, Cognitive Technologies, pages 13–36. Springer, Cham, Switzerland.

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

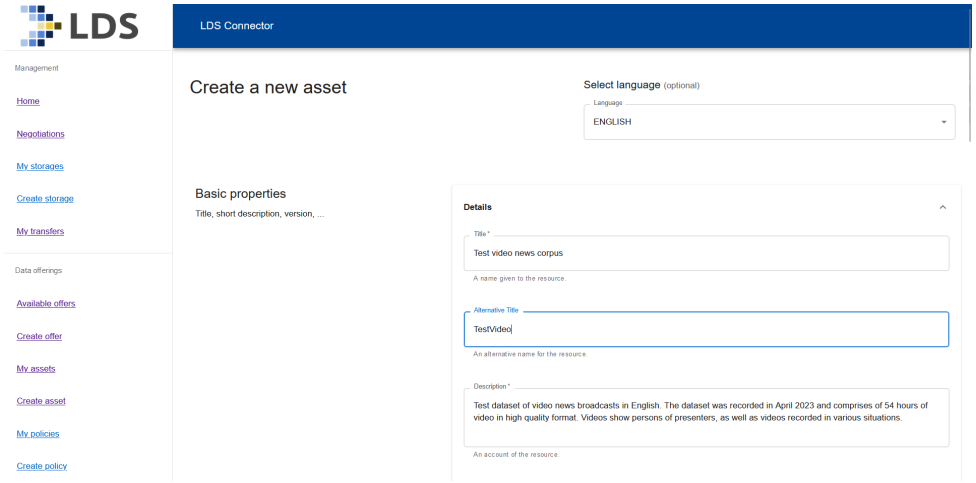
Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. 2021. European Language Grid: A Joint Platform for the European Language Technology Community. In *Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pages 221–230, Kyiv, Ukraine. ACL.

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 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turino, Italy. ELRA. Accepted for publication. May 20-25, 2024.

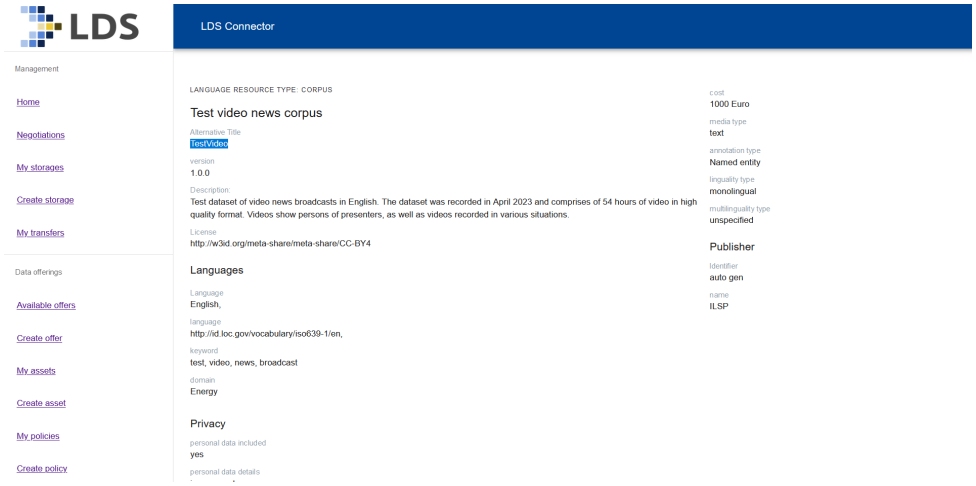
Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer.

A. Appendix

(1) Editing the metadata of a new data asset:



(2) Viewing a data asset:



(3) Creating a new offer:

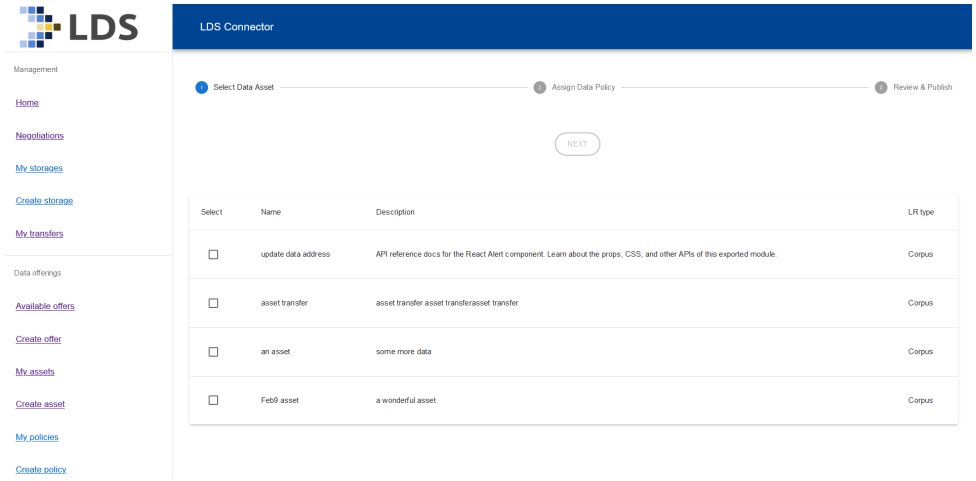
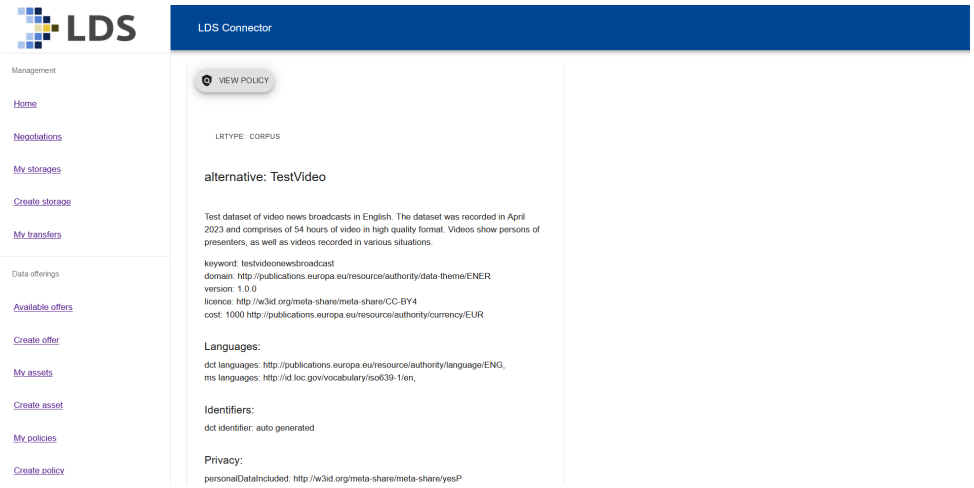
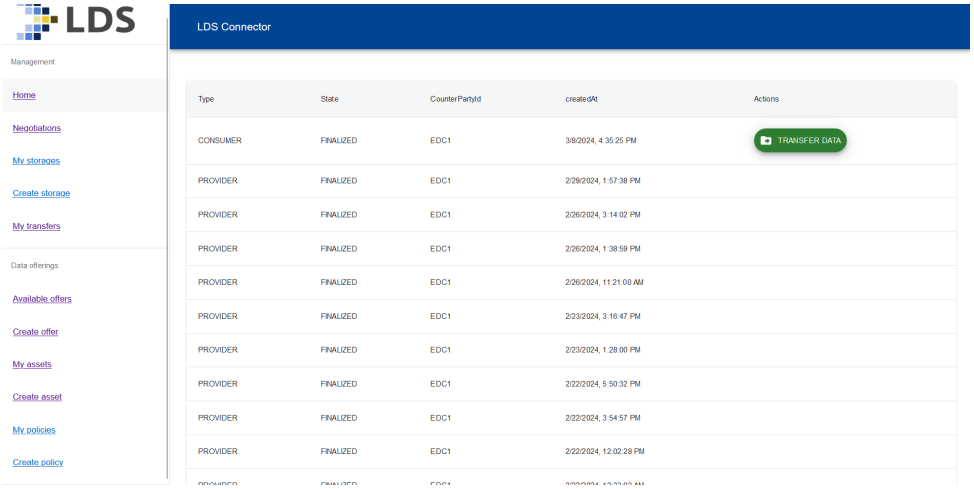


Figure 3: From the creation of a data asset to a list of transactions 1/2

(4) Viewing an offer:



(5) Listing all negotiations:



(6) Listing all transactions:

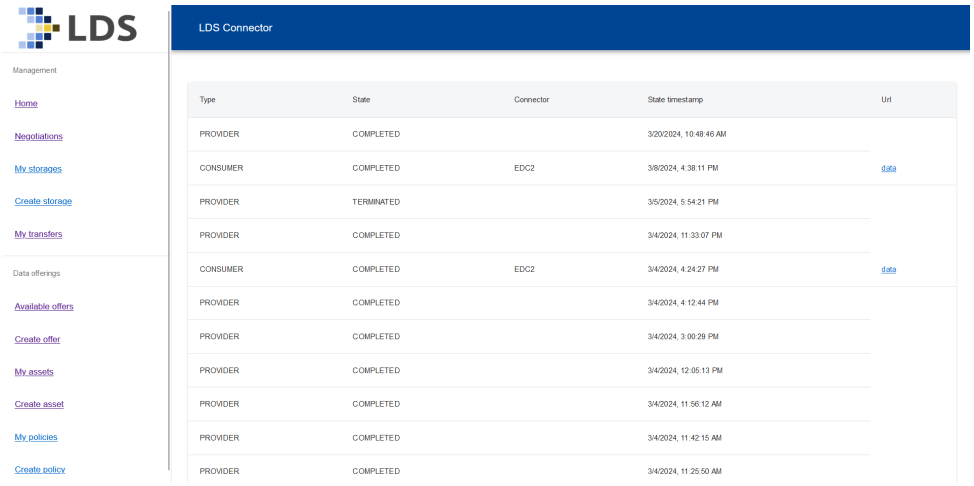


Figure 4: From the creation of a data asset to a list of transactions 2/2