WEBINAR

Data spaces: experience from the European Language Data Space

data.europa academy

11 April 2025

10.00 – 11.00 CEST

Rules of the game

The webinar will be recorded and published on the data.europa academy

For questions, please use the ClickMeeting chat

Please reserve 3 min after the webinar to help us improve by filling in our feedback form

data europa academy

Today’s speakers

Flora Kopelou
data.europa.eu,
Publications Office of the EU

Philippe Gelin
Head of Sector
Multilingualism,
DG CNECT

Stelios Piperidis
Senior researcher, Head
of the Institute for
Language and Speech
Processing, Athena RC

data europa academy

Agenda

10.00 – 10.05	Opening and introduction – *Flora Kopelou*
10.05 – 10.45	Deep-dive into the European Language Data Space – *Philippe Gelin, Stelios Piperidis*
10.45 – 11.00	Q&A session and closing remarks

data europa academy

The Language Data Space (and the ALT-EDIC)

Data.europa academy

Philippe GELIN
DG CONNECT – G3

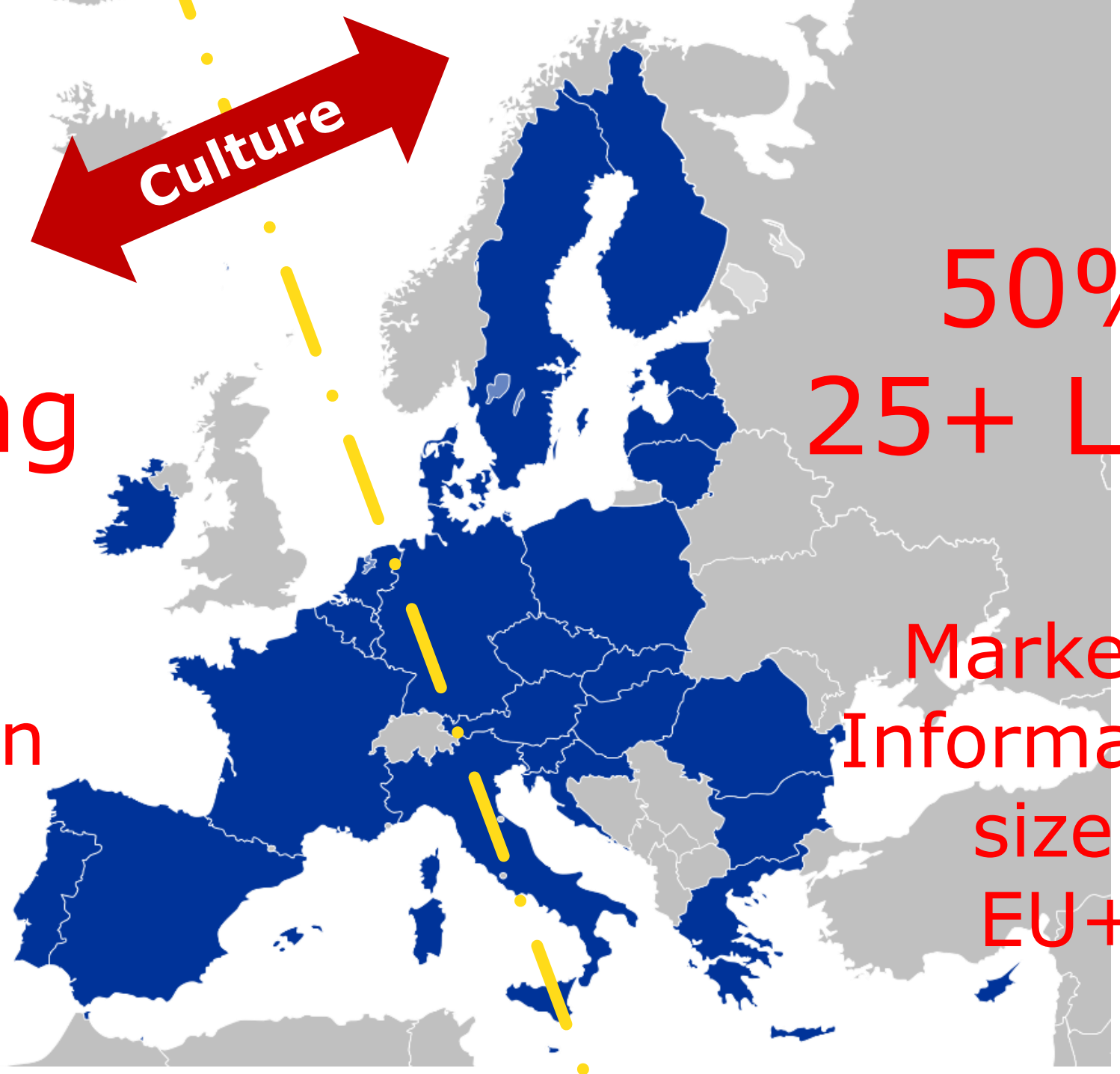
50%
English

Market /
Information
size:
EU/2

50%
24 Lang.

Market /
Information
size:
EU/24

European Commission



50%
25+ Lang

Market / Information size:
EU+

50%
25+ Lang

Market / Information size:
EU+

European Commission

LANGUAGE TECHNOLOGY MARKET STUDY

Impact of language technologies

Impact on intra-EU trade

The potential impact of language technologies on intra-EU trade can be in the order of EUR 360 billion.

99% of EU's companies are SMEs, adding 53% of value created.

For SMEs, the language divide and localisation costs are a high barrier to intra-EU trade.

Impact on productivity

The EU has a EUR 700 billion - 1.4 trillion economic opportunity with large language models.

Based on McKinsey, as newly added value and labour productivity gains.

The big underlying assumption is that LLMs support all languages at the same level.

LTs as costs are insignificant when compared to opportunities.
Especially when LTs are being commoditised.

LANGUAGE TECHNOLOGY MARKET STUDY

Defining language technology

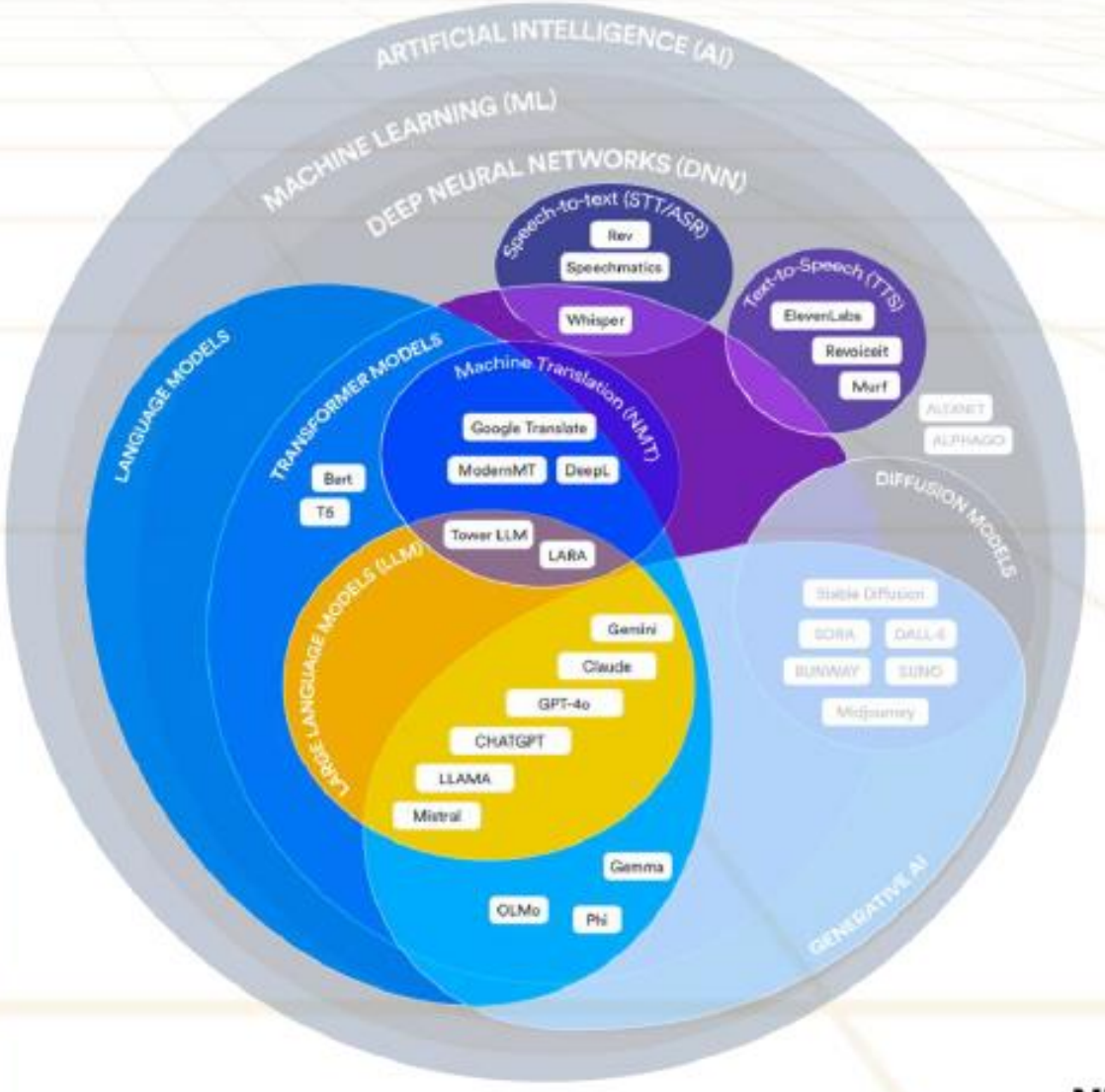
AI is eating the world of (language) technology

LTs develop rapidly, and so does the market and the expectations.

Progress is measured in weeks rather than months and years.

Investment is massive, revenues are lagging.

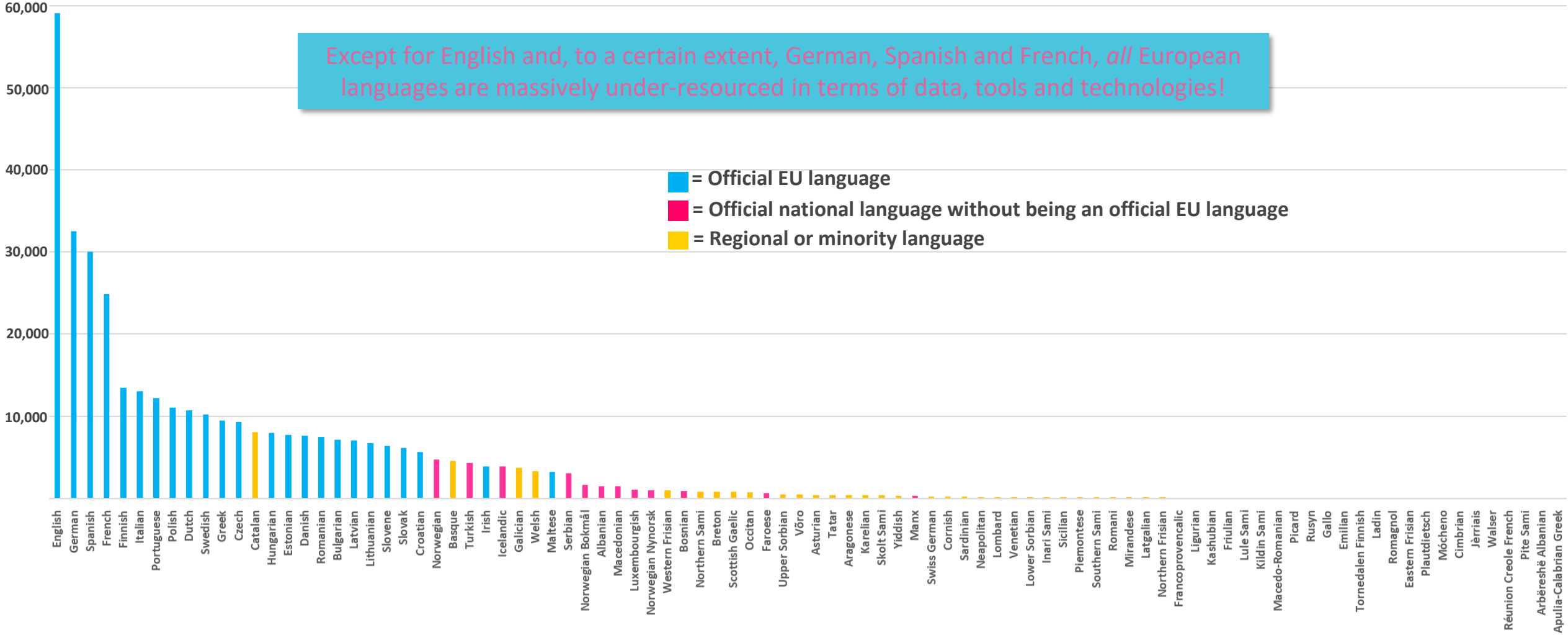
One key limitation is the vast amounts of high-quality data needed for model (re-)training.



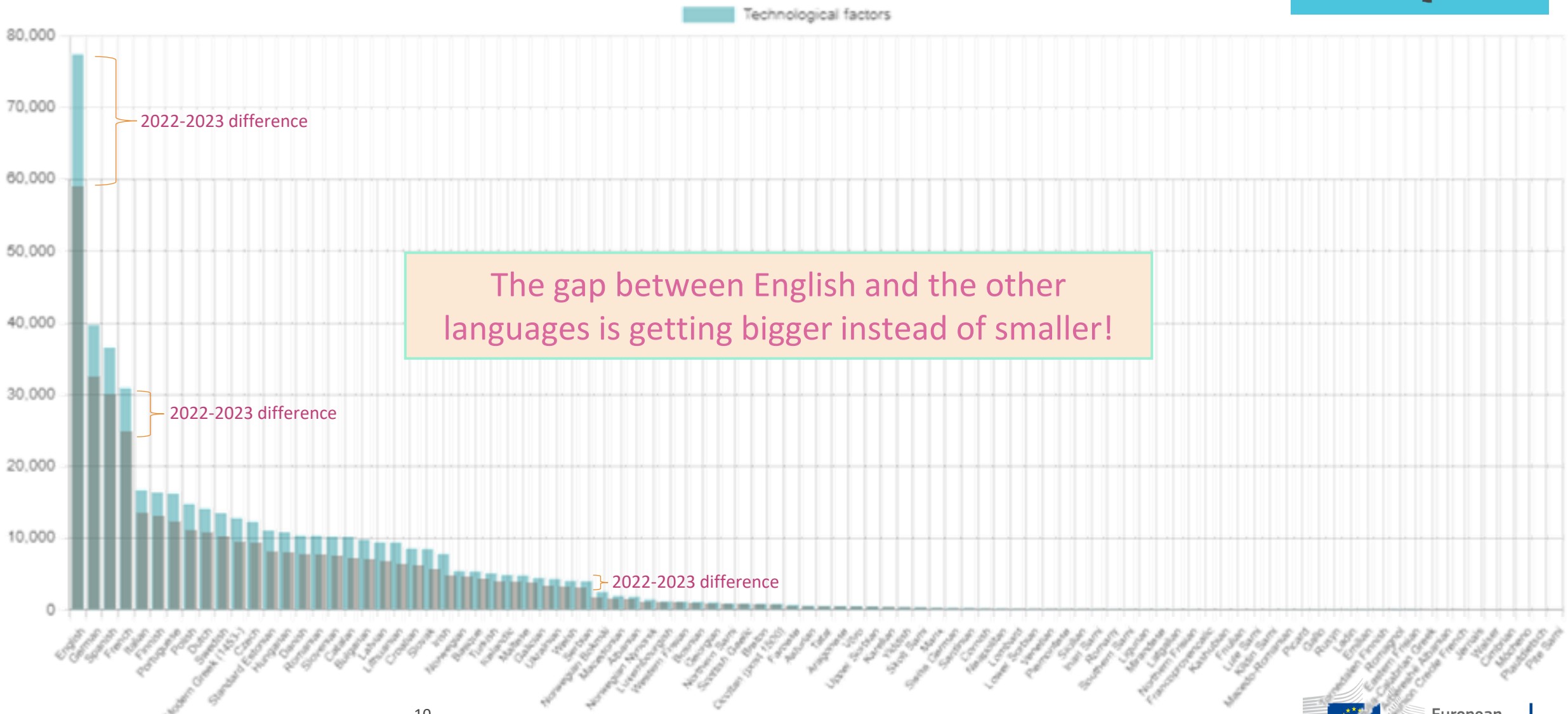
Source: The Language AI Alphabet, Nimdzi.com (2024)
https://www.nimdzi.com/the-language-ai-alphabet-transformers-llms-generative-ai-and-chatgpt/

Nimdzi

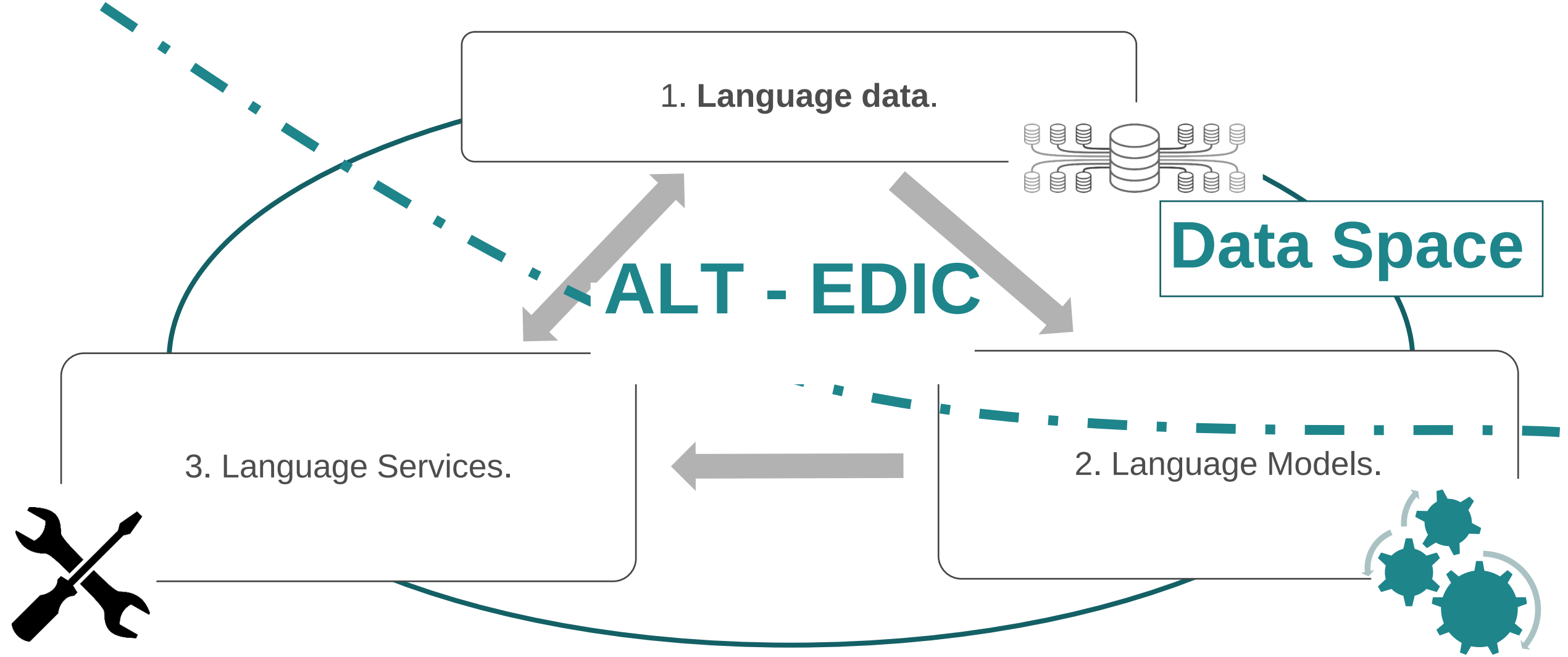
Digital Language Equality Metric: Technological Scores



DLE Metric: 2022 vs. 2023



European Language Technologies Landscape



alt-edic

Alliance for Language Technologies European Digital Infrastructure Consortium

Objectives

- Preserve linguistic and cultural diversity in Europe
- Technological leadership and strategic autonomy
- Respect European rules and values
- Cooperation
- Raising awareness

Participation

Members [18]

FR, BG, CZ, DK, ES, FI, GR, HR, HU, IE, IT, LT, LV, LU, NL, PL, SI, Flanders

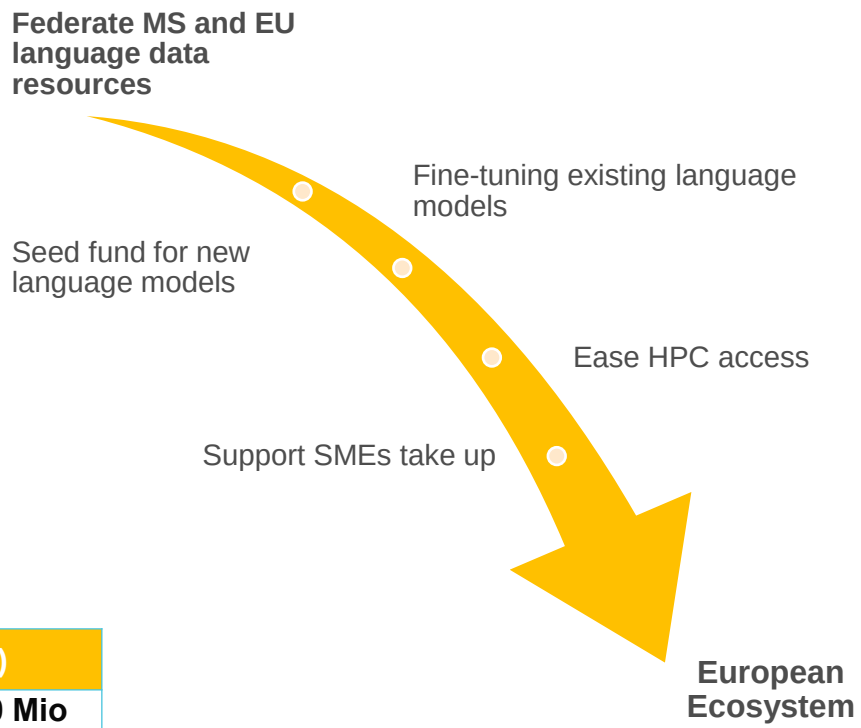
Observers [8]

AT, BE, CY, EE, MT, PT, RO, SK

Interested [2]

DE, SE

Actions Plan



Budget

Projects (Costs)	
- LLMs4EU	€ 40 Mio
- ALT-EDIC4EU	€ 4 Mio
- OpenEuroLLM	€ 40 Mio
- LLM-BRIDGE	€ 6,8 Mio

European Commission

Common European Language Data Space

- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€ if renewed) – runtime: 36 months (+ 12 months if renewed)
- Objective: Develop and deploy a trustworthy and secure European infrastructure and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- LDS is a marketplace for all organisations – commercial and public.
- LDS provides helpdesks for legal and for technical questions.
- Salient features: governance framework, technical infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens
- The four core partners have been involved in many projects.
- Technical development informed by ELG, ELRC-SHARE, META-SHARE.

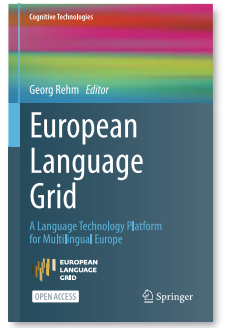
Lead Partner and Coordinator
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Partners and Operation Leads
R.C. "Athena", Institute for Language and Speech Processing
Evaluations and Language Resources Distribution Agency
TILDE
Main Subcontractors
3pc GmbH Neue Kommunikation
CLARIN ERIC
Big Data Value Association (Data, AI and Robotics) AISBL

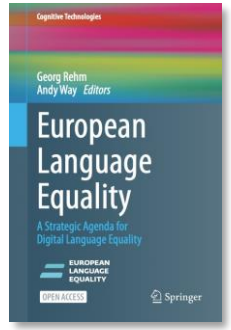
META-NET

European Language Resource Coordination

EUROPEAN LANGUAGE GRID

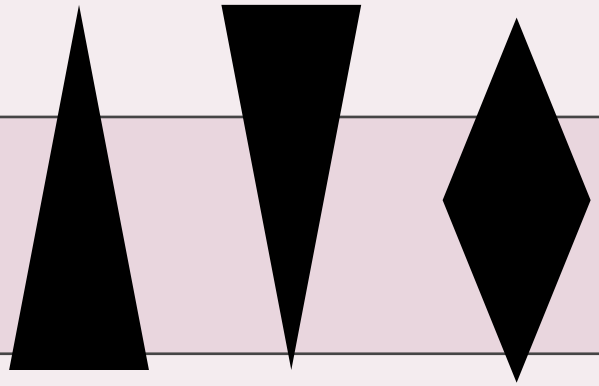
EUROPEAN LANGUAGE EQUALITY





Classes of Data

Class of Data	Typical Size	Providers	Integration into LDS	Relevance, especially for LLM Development	Price per word	Market size	Expected Interest
Web Crawls	Very big (TB, PB)	A handful of. (Common Crawl, Internet Archive, Paracrawl, … etc)	Challenge due to their size (hard to transfer, hard to preprocess, hard to store; must be close to the HPC)	LLM training purpose. Indispensable due to their size and coverage – but high level of noise, massive need for pre-processing.			
Regular Corpora and Language Resources	Small (MB, GB)	Primarily NLP/LT research: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Can be easily integrated by connecting the repositories to LDS	LLLM training / adaptation. Usually very high-quality data and thus relevant for LLMs but not as base data			
New, fresh data from industry and other organisations	Arbitrary size, ideally as large as possible	Public administration, Publishing houses, media companies, libraries, call centres, broadcasters etc.; also: Media Data Space	Can be easily integrated by connecting these organisations to LDS	LLM adaptation / fine-tuning: High quality data; domain-specific data; data covering specific languages; raw data; processed data; evaluation data; data relevant for LLM adaptation. Easily aggregated.			
Data from Individual companies (sensitive)	Small	Large bases (SMEs)	Easily integrate, but potential higher maintenance cost vs benefit	LLM fine-tuning: High quality, relevant for fine-tuning LLMs. Limited usage for focussed sectors. (potential for aggregation / data brokers)			



Language Data Space Marketplace – Value Proposition

Data Providers

- Sell data products (= data sets, data offers)
  - Find new customers
  - Extend or enrich datasets using AI/NLP services offered in the wider LDS ecosystem
- Legal compliance by design
  - Stay in control over use and access of data
  - Compliance with EU regulation and standards
- Limited effort
  - Keep existing infrastructure and workflows
  - Interoperability with other data spaces
  - Legal and technical helpdesks available
- Contribute to European LLMs: *from* and *for* Europe

Data Consumers

- Buy or access data products to develop better AI-based services (including LLMs)
  - Multilingual data
  - Multimodal data
  - Domain-specific data
  - All European languages
  - Easy discoverability and access
- Limited effort: keep existing infrastructure
- Legal compliance by design
  - Compliance with EU regulation and standards
  - Transparency: emphasis on data provenance
- Find new customers for services and products

Technical Helpdesk – Legal Helpdesk

- **Legal helpdesk:** provides legal advice and assistance to all stakeholders involved in the LDS. It provides guidance for legal questions related to the collection and sharing of language data, licensing schemes, IPR clearance, data protection requirements as well as confidentiality aspects. These may include:
  - questions related to any kind of utilisation of language data in the LDS;
  - legal queries, e.g., how to exploit the LDS by clearing specific legal aspects, etc.

- **Technical helpdesk:** provides technical advice and assistance to all stakeholders involved in the LDS, providing first-line support to technical questions. These may include:
  - general troubleshooting and assistance on the usage of the LDS;
  - technical support, e.g., validation of technical compliance of data assets;

EVENT | Publication 19 March 2025

Language Technology Landscape Conference 2025

15 April 2025

Online | 15 April | 9:00 - 12:00

The Language Technology Landscape Conference disseminates the findings from the Market Study conducted by Nimdzi Insights.

Nimdzi Insights and its partners are conducting a comprehensive Market Study on Language Technologies for the European Commission to assess the development, quality, and global impact of language technologies globally and in Europe. The Language Technology Landscape Conference

Organiser

EUROPEAN
LANGUAGE
DATA SPACE

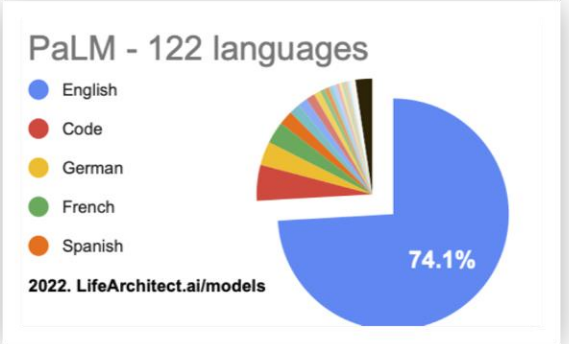
The European Language Data Space
Principles, Architecture and Components

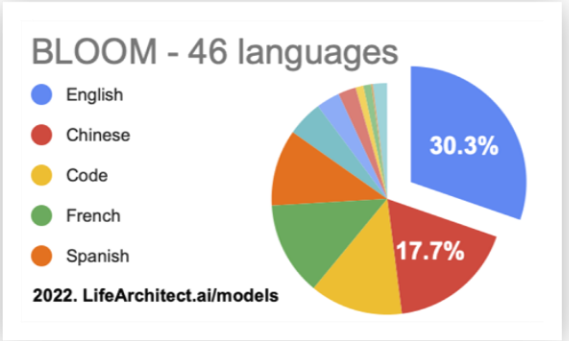
Stelios Piperidis (Athena RC, Greece)
spip@athenarc.gr

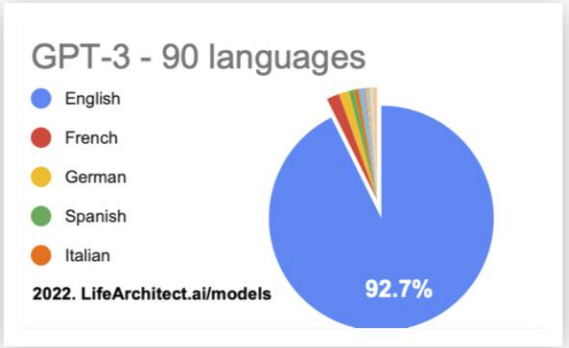
11-04-2025 Data spaces: experience from the European Language Data Space
https://language-data-space.ec.europa.eu

Artificial Intelligence – Large Language Models (LLMs)

- Unprecedented capabilities: Large language models are the most disruptive breakthrough in AI in recent history (GPT-3, ChatGPT, GPT-4, Claude, Gemini, Llama, DeepSeek etc.)
- LLMs are trained on vast amounts of data and optionally also image, video, audio etc. data, i.e., multimodal data
- Multilingualism makes everything much harder (data imbalance): Europe's languages are vastly under-resourced, except English
- Unprecedented opportunities:
  - The global LT/NLP market is expected to reach 439.85B$ by 2030
  - The global Gen AI market is expected to reach 1.3T$ by 2032
- A concerted effort for the collection of data for all European languages is very much needed to be able to develop LLMs according to our needs and cultures …
- … and to make a difference with European data and European stakeholders.
- Already now billions and billions are being invested



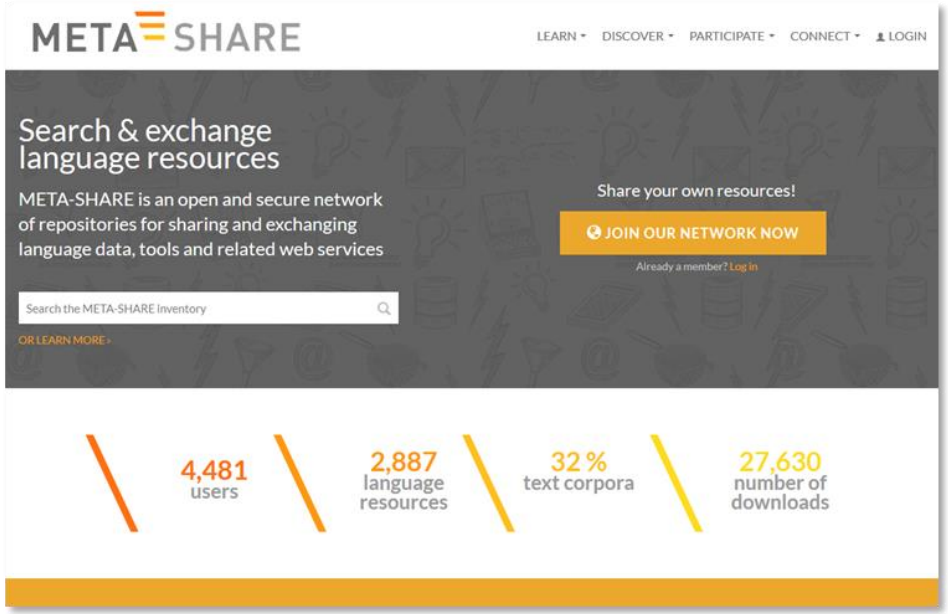


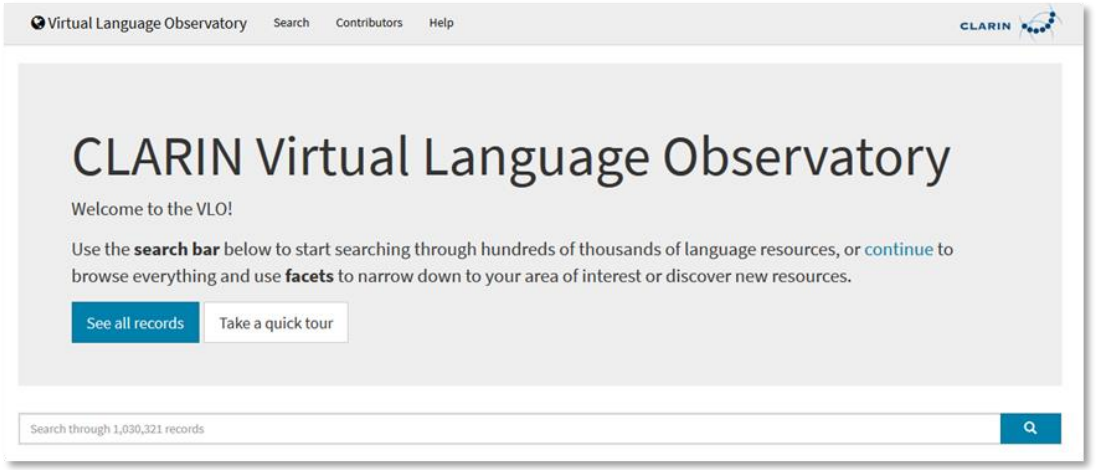


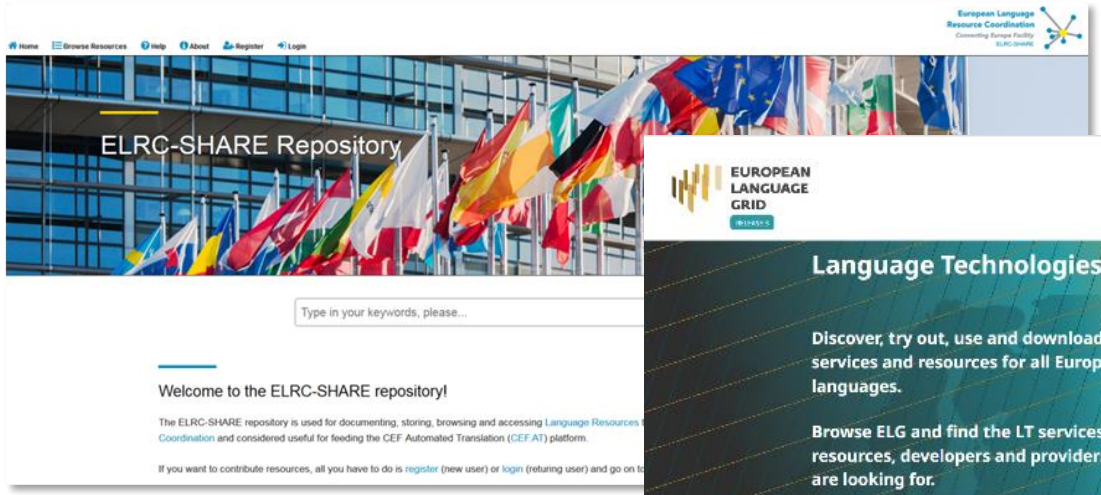
European Initiatives

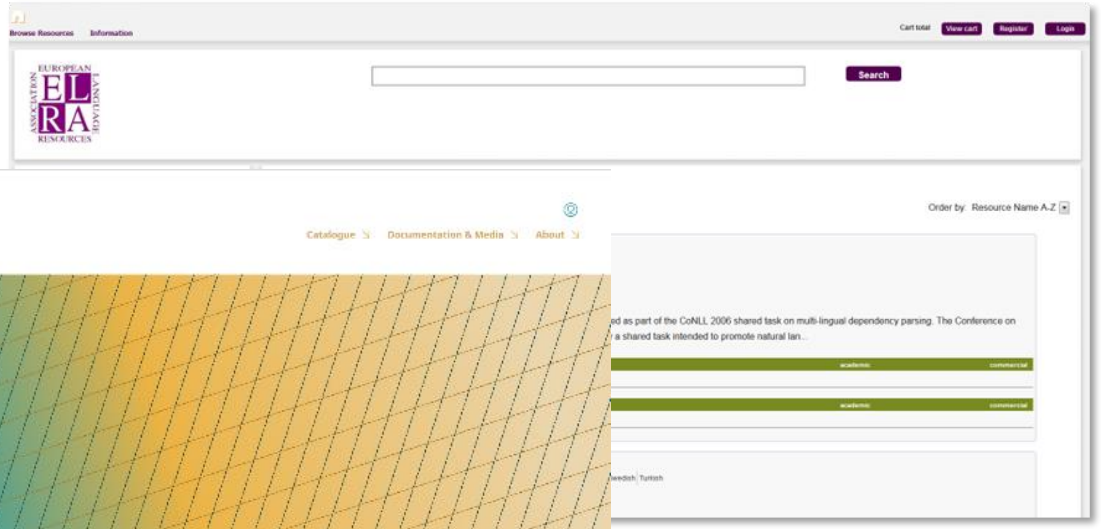
- European initiatives for the development of LLMs
  - Large research projects in almost every country, e.g., Spain, Denmark, Italy, Germany etc.
  - Companies in many countries, e.g., Finland (Silo.ai), France (Mistral), Germany (Aleph Alpha)
  - EU-funded projects, e.g., HPLT, TrustLLM, OpenEuroLLM, LLMs4EU
  - New pan-European initiative: ALT-EDIC
  - New EU initiative: AI Factories (tightly coupled with national HPC centres and EuroHPC JU)
- Challenges:
  - HPC facilities (amount, access and ease-of-use)
  - Speed of the big tech players in the US and Asia vs. speed of Europe
  - Availability of data for *all* European languages – *right now the most crucial bottleneck by far*

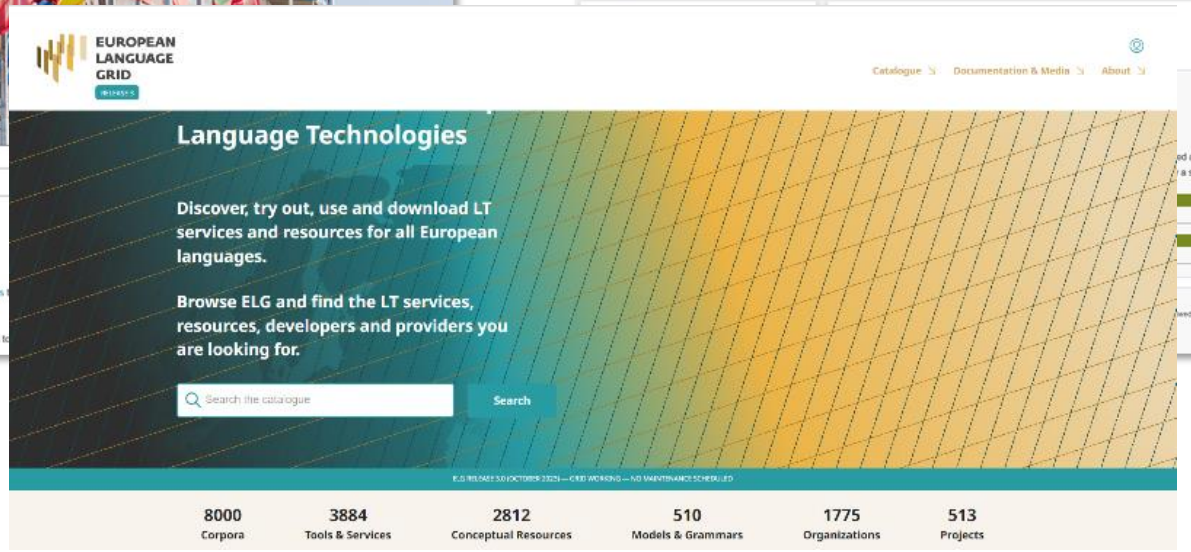
Long History of Language Data Sharing











Build on existing achievements

Respond to language data requirements for multilingual Europe in the AI era

Facilitate the language data economy

Common European Language Data Space

EUROPEAN LANGUAGE DATA SPACE

- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€ if renewed) – runtime: 36 months (+ 12 months if renewed)
- Objective: Develop and deploy a trustworthy and secure European infrastructure and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- LDS is a marketplace for all organisations – commercial and public
- LDS provides helpdesks for legal and for technical questions
- Salient features: governance framework, technical infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens
- The four core partners have been involved in many projects.
- Technical development informed by ELG, ELRC-SHARE, META-SHARE.

Lead Partner and Coordinator
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Partners and Operation Leads
R.C. “Athena”, Institute for Language and Speech Processing
Evaluations and Language Resources Distribution Agency
TILDE
Main Subcontractors
3pc GmbH Neue Kommunikation
CLARIN ERIC
Big Data Value Association (Data, AI and Robotics) AISBL

LDS contributing to the Data Spaces ecosystem

data sovereignty

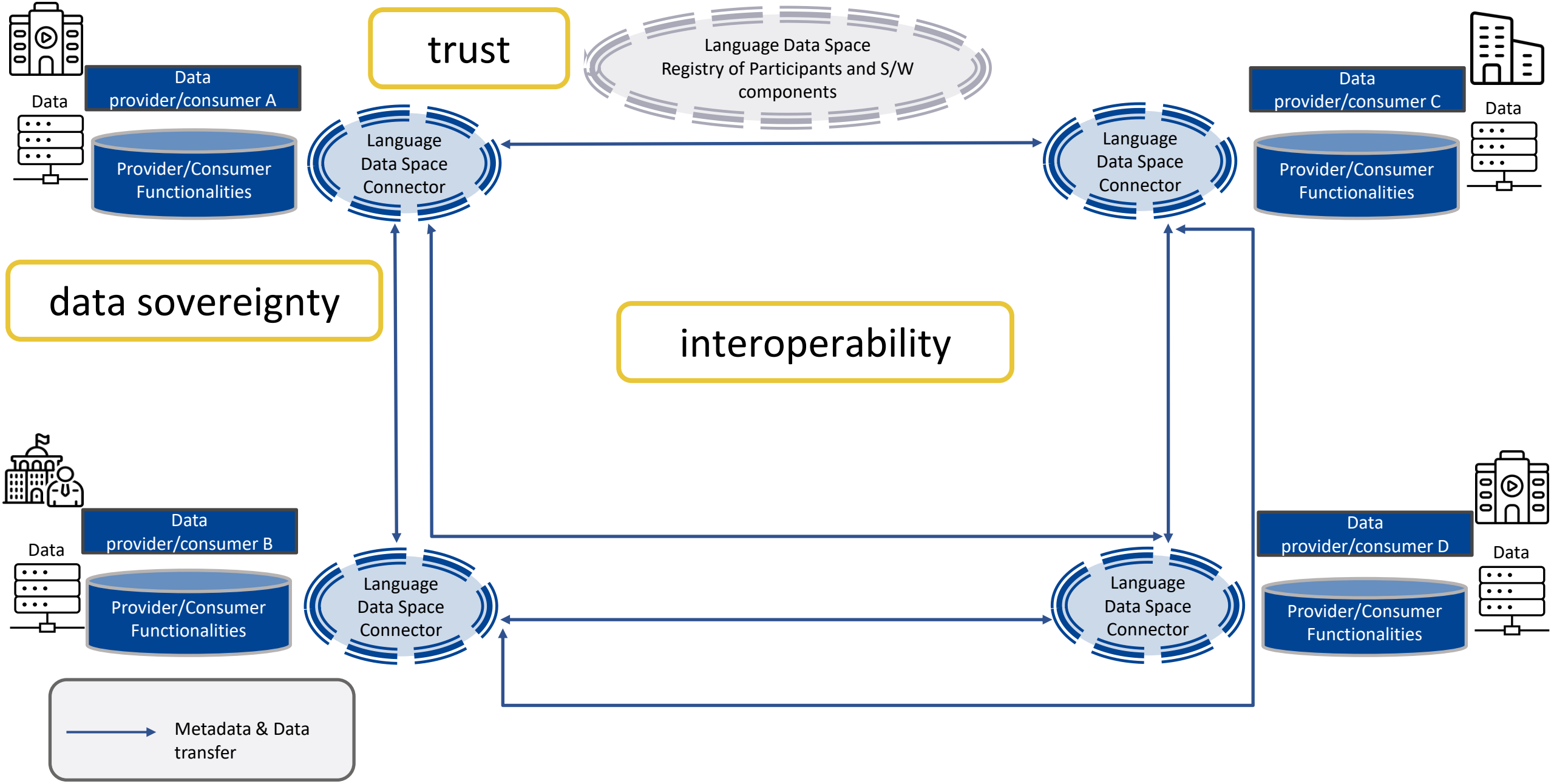
trust

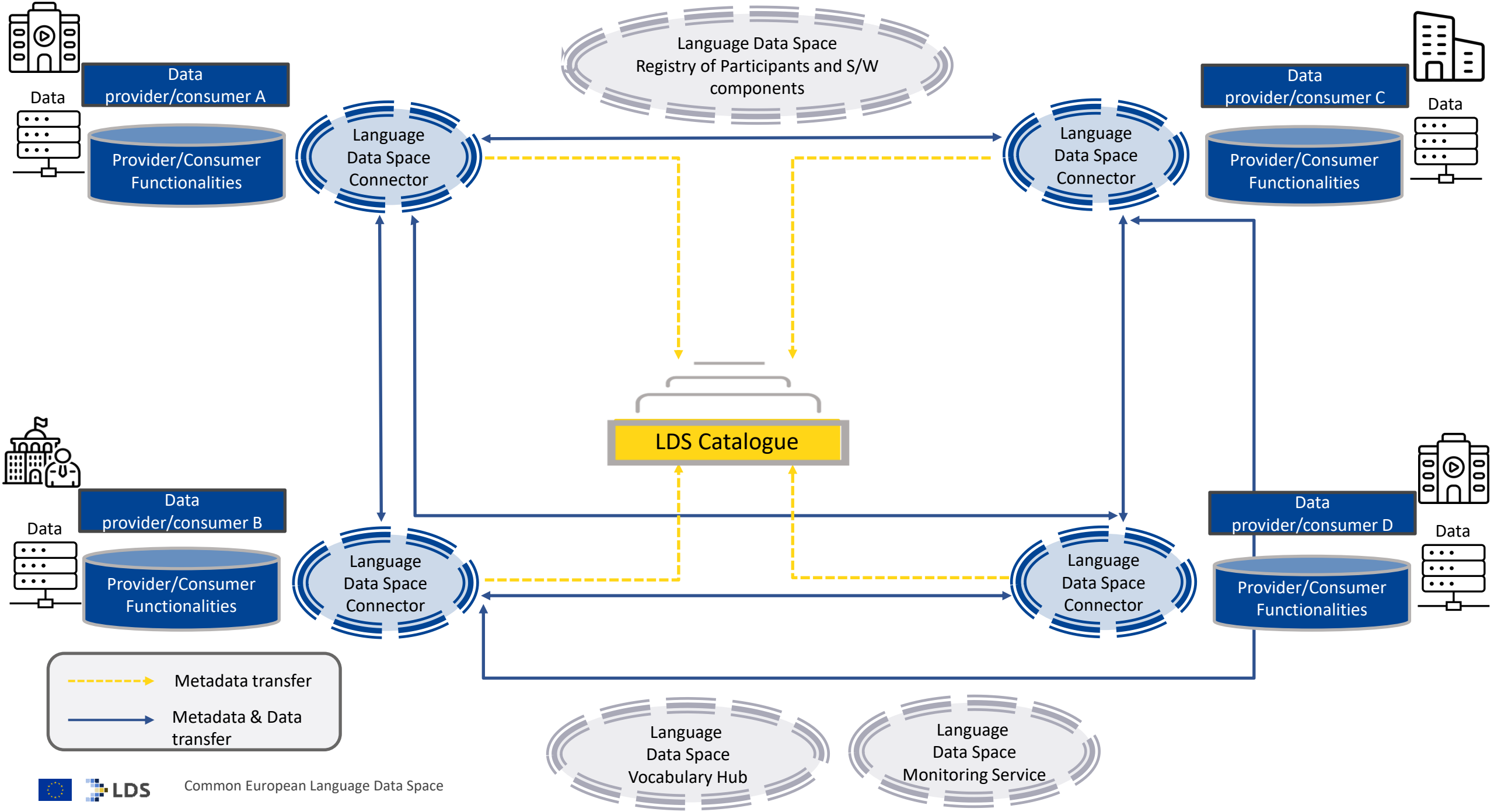
interoperability

LDS Architecture

- Objective 1: **provide technological support for complying with the Data Spaces principles**
  - sovereignty, trust, interoperability

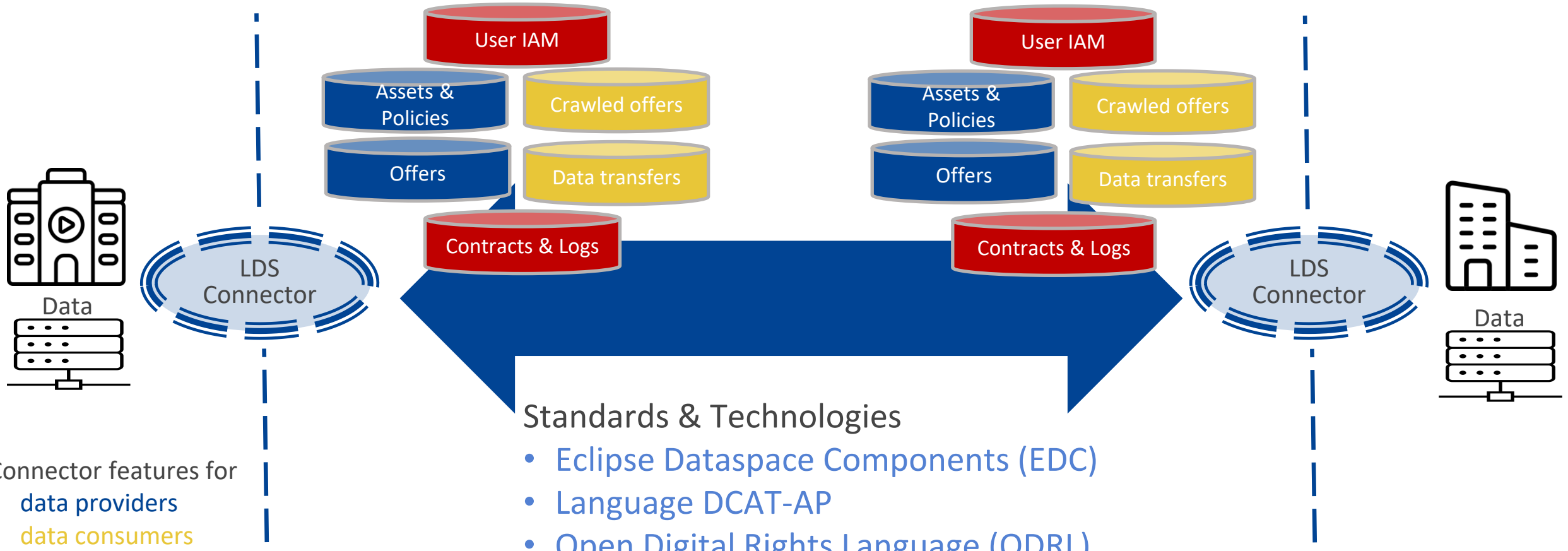
- Objective 2: **provide technological support for LDS workflows** designed
  - for data providers & consumers
  - for the LDS Governance Board


trust
Language Data Space
Registry of Participants and S/W components
Data
Data provider/consumer A
Provider/Consumer Functionalities
Language Data Space Connector
Language Data Space Connector
Data provider/consumer C
Data
Provider/Consumer Functionalities
data sovereignty
interoperability
Data
Data provider/consumer B
Provider/Consumer Functionalities
Language Data Space Connector
Language Data Space Connector
Data provider/consumer D
Data
Provider/Consumer Functionalities
Metadata & Data transfer


Language Data Space
Registry of Participants and S/W components
Data
Data provider/consumer A
Provider/Consumer Functionalities
Language Data Space Connector
Language Data Space Connector
Data provider/consumer C
Data
Provider/Consumer Functionalities
LDS Catalogue
Data
Data provider/consumer B
Provider/Consumer Functionalities
Language Data Space Connector
Language Data Space Connector
Data provider/consumer D
Data
Provider/Consumer Functionalities
Metadata transfer
Metadata & Data transfer
Language Data Space Vocabulary Hub
Language Data Space Monitoring Service

LDS Connector

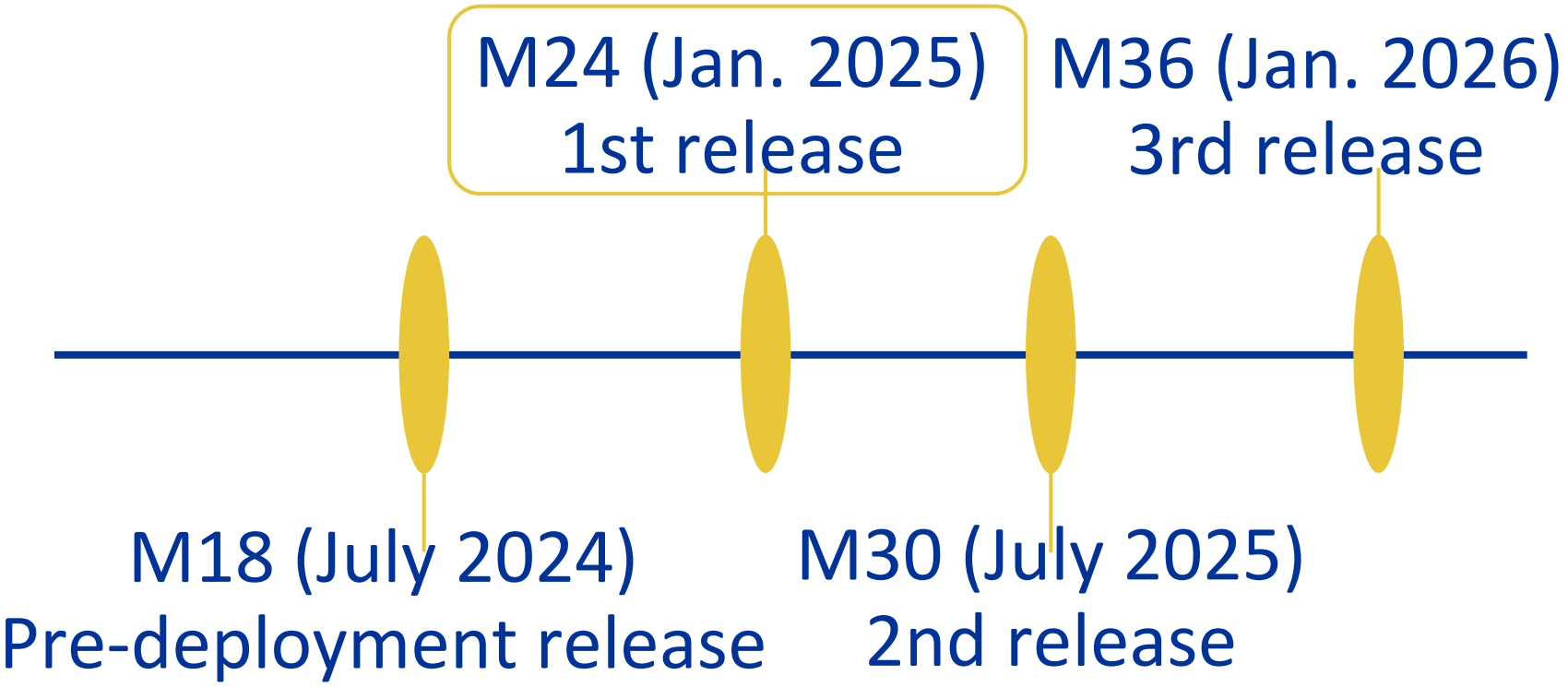
- Asset: Data/Files+Metadata
- Offer: Asset+Policy



Connector features for
- data providers
- data consumers
- data prosumers

Standards & Technologies
- Eclipse Dataspace Components (EDC)
- Language DCAT-AP
- Open Digital Rights Language (ODRL)
- Dataspace Protocol (DSP)

LDS Release Plan



v1.0.0-beta

LDS v1.0.0-beta

1. LDS participant application and onboarding

- Organisation's **legal representative applies** online (https://ldssetup.ilsp.gr) and accepts the LDS Terms of Service
- The data are added in the **LDS Registry of participants**
- **LDS Governance Board** assesses the application (to ensure that it complies with the eligibility criteria) ➔ approves or rejects (providing the reasons)
- **Upon approval,** participant downloads, installs & configures the **LDS connector at their own infrastructure**

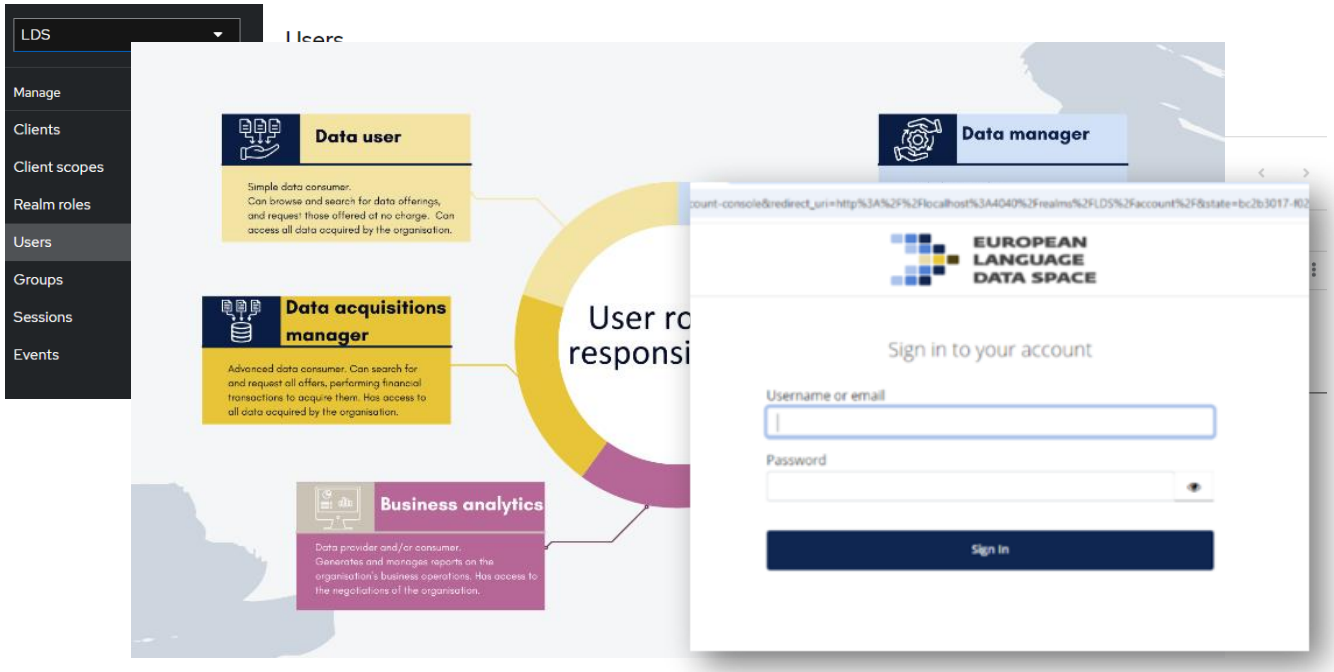
➔ Issuance of **authentication keys**

- **Registers** the **Connector** at the LDS Registry
- Configures their **storage solutions**

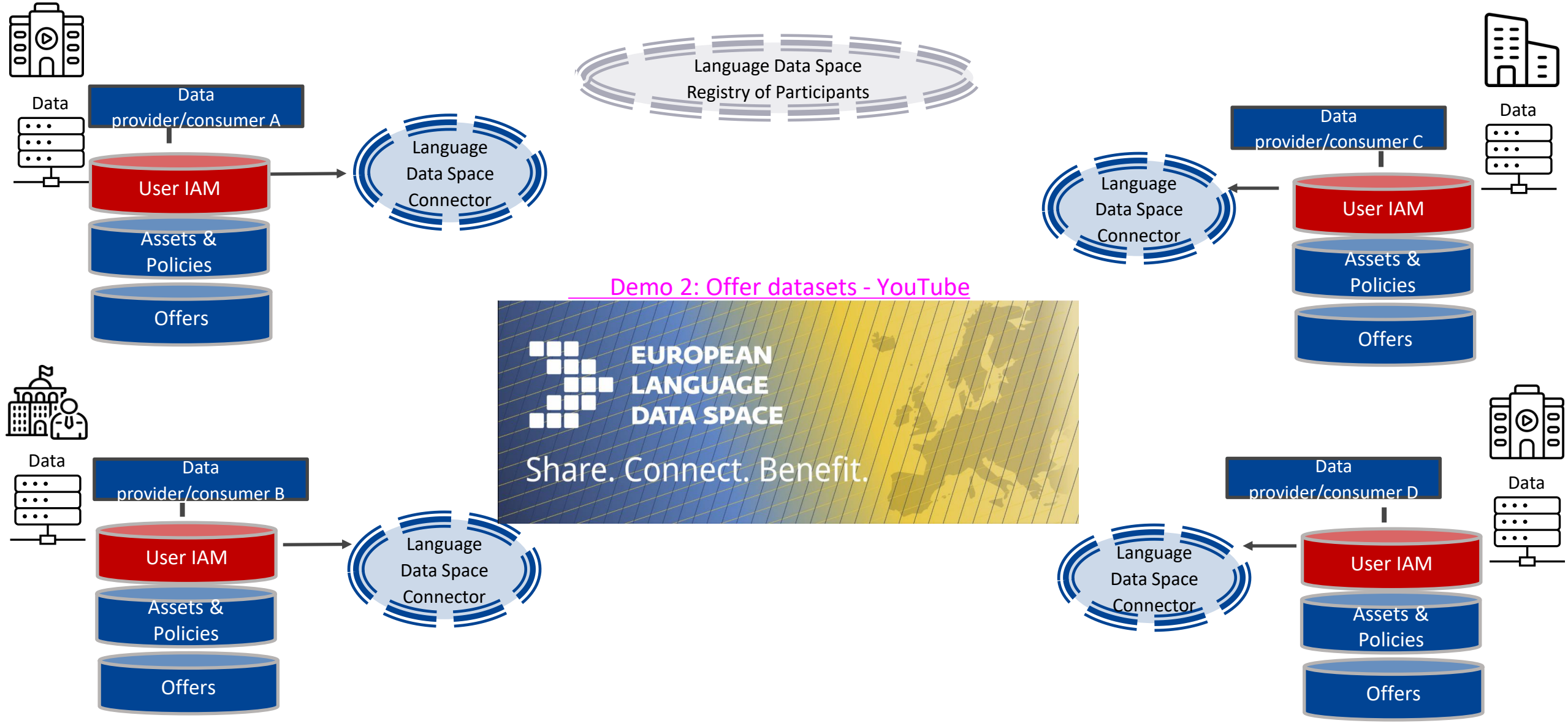
Demo 1: Apply for a membership in the LDS - YouTube

2. User Identity Access Management (inside the participant's LDS connector)

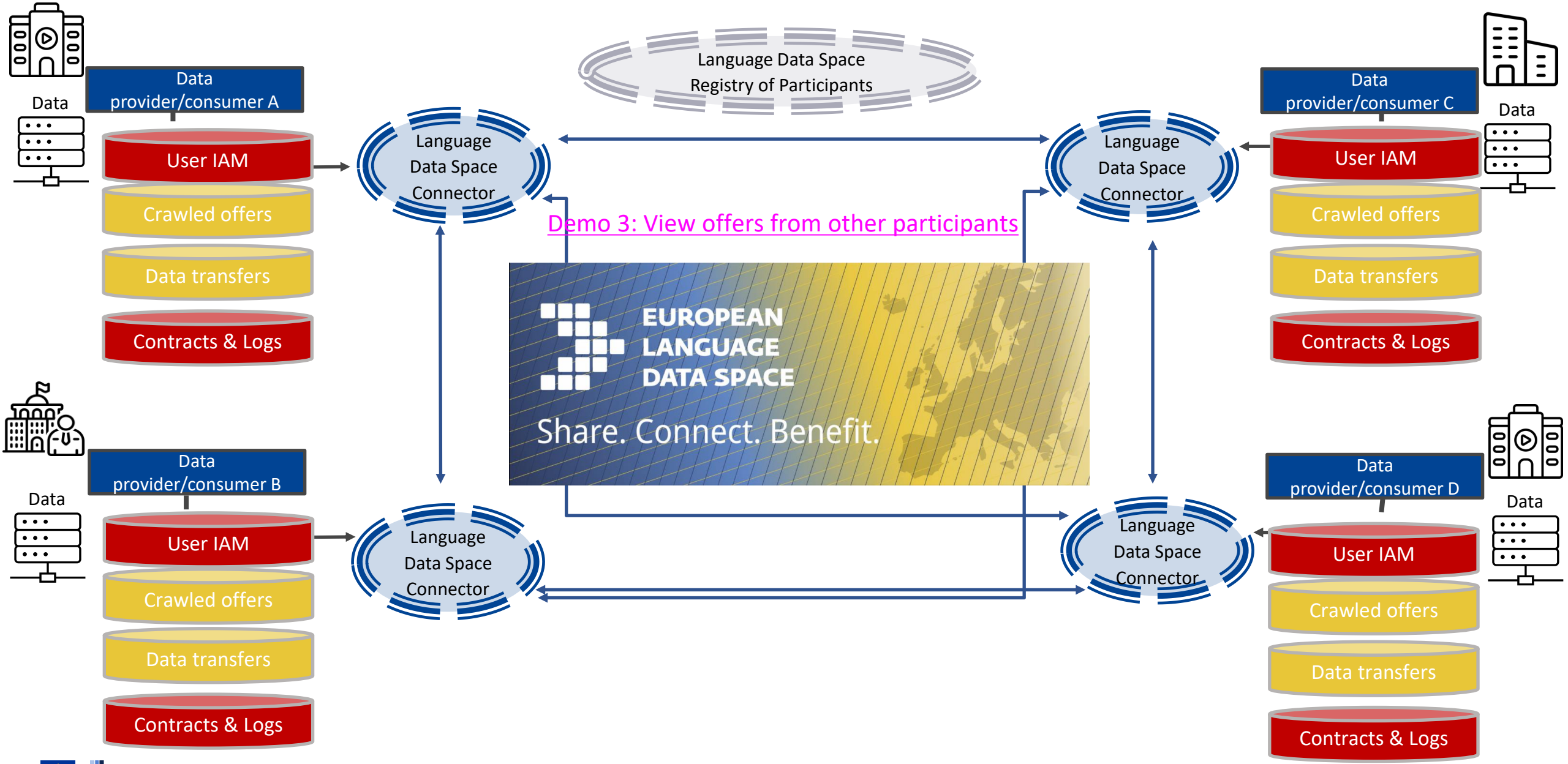
- **Keycloak solution**
- **Registration** of users by user administrator (with possibility to use active directory)
- Assignment of pre-defined LDS **user roles** for providers & consumers with simple and advanced rights (with possibility to assign multiple roles)



3. Data Providers - Creating assets, policies, and offers

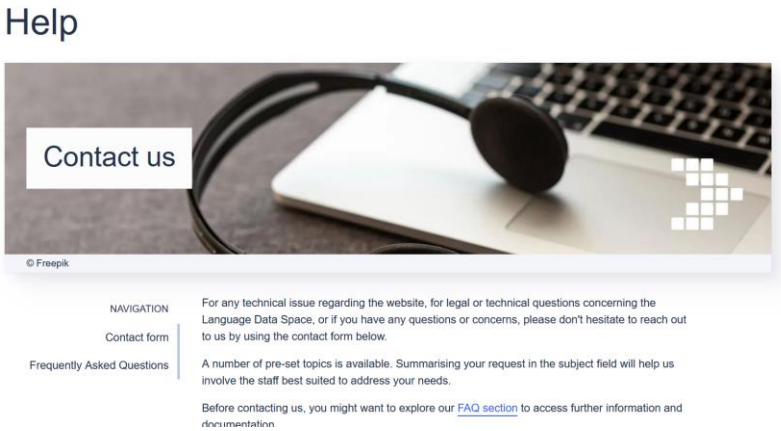


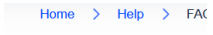
4. Data Consumers – offers discovery, negotiation, contracts, download



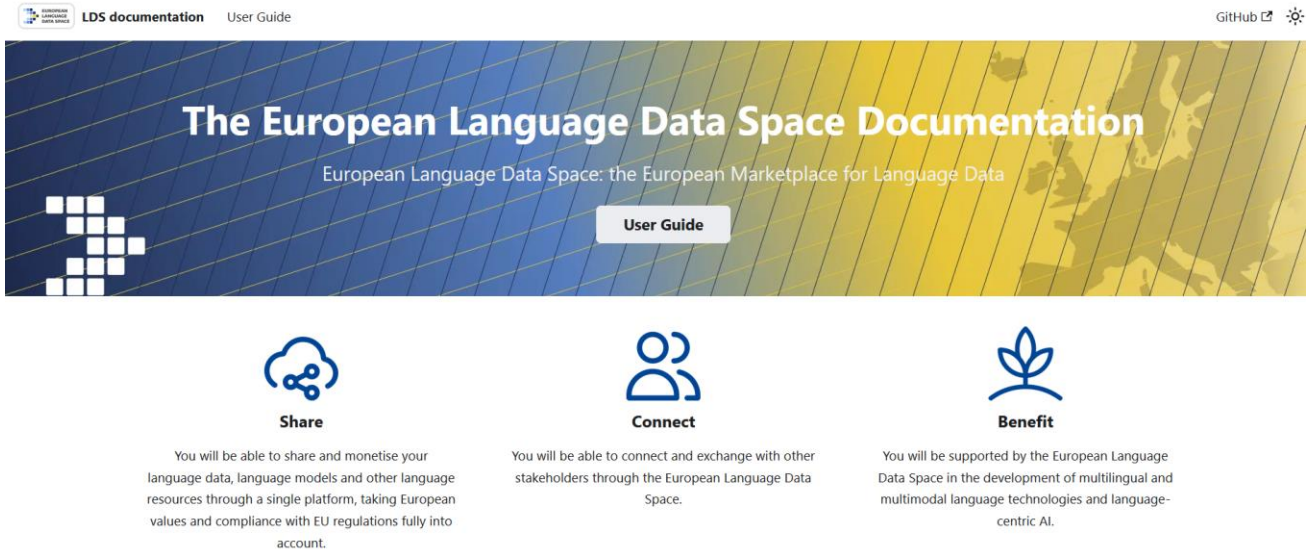
Technical support

- LDS contact https://language-data-space.ec.europa.eu/help_en
- FAQs https://language-data-space.ec.europa.eu/help/faq_en
- LDS technical documentation
  - for technical administrators
  - for LDS users (providers & consumers)
  - for LDS Governance Board members





FAQ



Data Spaces Principles

Trust framework

- applicants assessed by LDS Governance Board
- applicants must consent to LDS Terms of Service
- transactions only between authenticated & authorised connectors
- IAM for individuals inside organisation
- data transfer only upon conclusion of contract (between known contracting parties)

Data sovereignty

- data remain at one's own storage space
- participants define their data policies
- policy enforcement before approval of data request

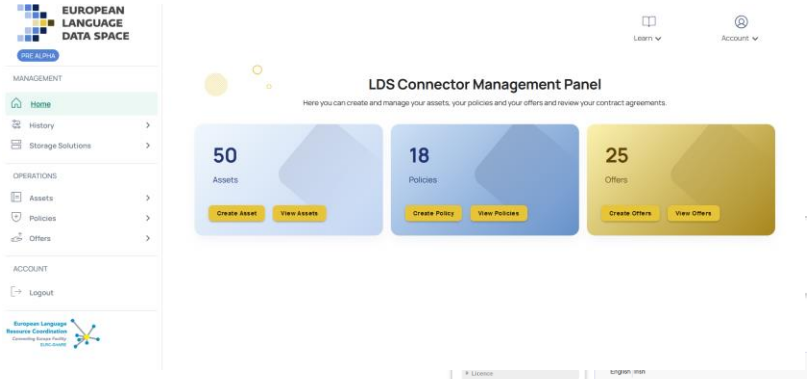
Provenance & traceability

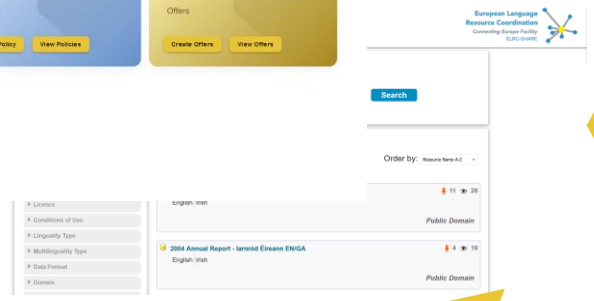
- contracts stored at both contracting parties' connectors
- logging & monitoring of transactions locally & centrally

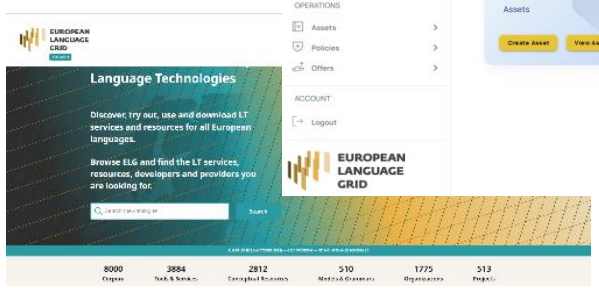
Interoperability

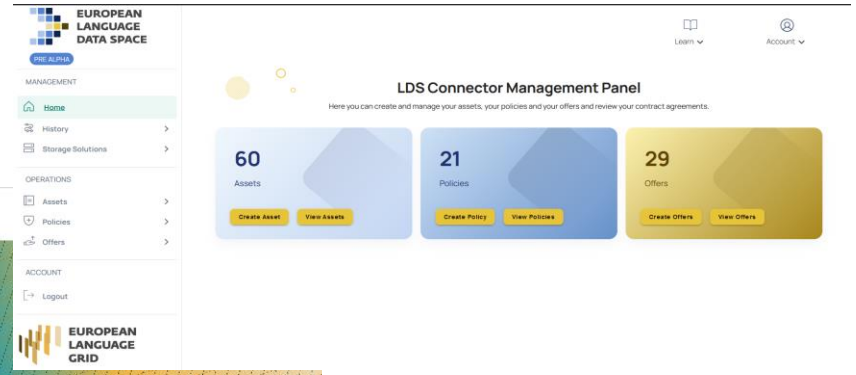
- compliant with Dataspace Protocol
- adoption of standard transfer protocols for data transfer
- deployment of standard model for self-descriptions (based on DCAT model)
- deployment of ODRL model for policies
- recommendations for standard models & formats for data

Demo/Testing Setup





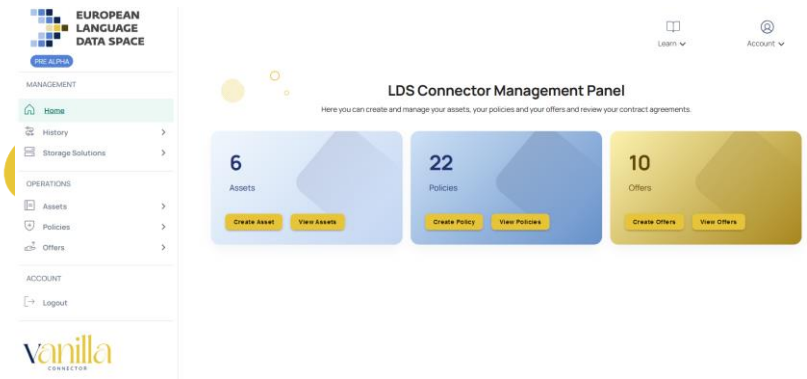




Apply at

https://ldssetup.ilsp.gr/ to

join LDS and test it



Classes of Data

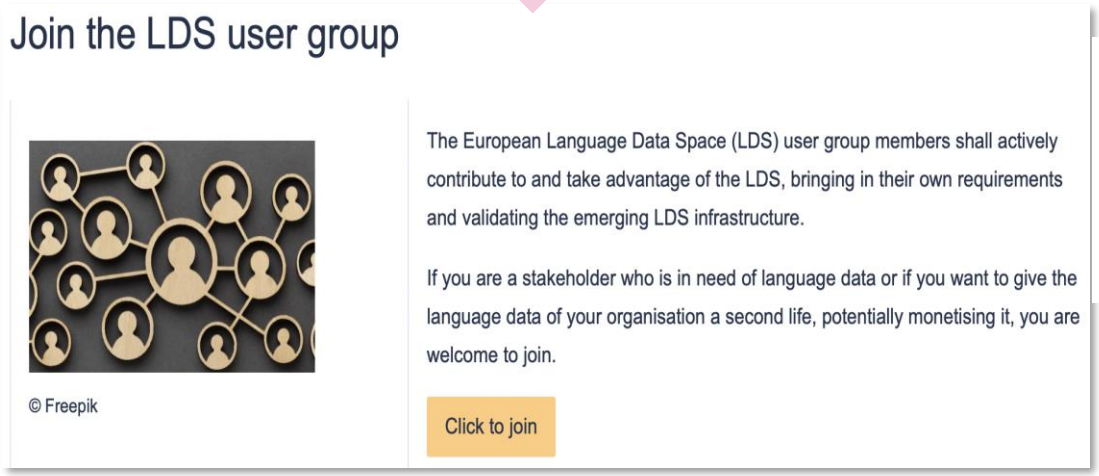
Class of Data	Typical Size	Providers	Integration into LDS	Relevance, especially for LLM Development
Regular Corpora and Language Resources	Small (MB, GB)	Primarily NLP/LT research: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Can be easily integrated by connecting the repositories to LDS	Usually very high quality data and thus relevant for LLMs but not as base data
Web Crawls	Very big (TB, PB)	Common Crawl (and OSCAR-processed CC dumps), Internet Archive dumps etc.	Challenge due to their size (hard to transfer, hard to preprocess, hard to store; must be close to the HPC)	Indispensable due to their size and coverage – but: high level of noise, massive need for pre-processing
New, fresh data from industry and other organisations	Arbitrary size, ideally as large as possible	Publishing houses, media companies, libraries, call centres, broadcasters etc.; also: Media Data Space	Can be easily integrated by connecting these organisations to LDS	High quality data; domain-specific data; data covering specific languages; raw data; processed data; evaluation data; data relevant for LLM development etc.

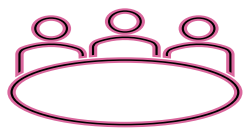
LDS User Group, LDS Focus groups

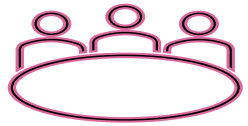
- Mailing list for the LDS user group
- The LDS user group is growing

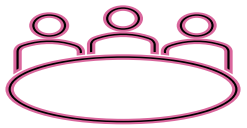
- **If you're interested, please get actively involved and join the LDS User Group!**
  - Validation of concepts, ideas, software; first test installations of the LDS connector taking place; first trial exchanges of data; surveys etc.
  - You can also help on a more substantial, in-depth level by joining the **LDS Focus Groups on the LDS Connector, Metadata & Data and Policies** – please approach us if you are interested.

https://language-data-space.ec.europa.eu




LDS Connector


Metadata & Data


Policies

Common European Language Data Space

Thank you!

Stelios Piperidis (Athena RC, Greece)
spip@athenarc.gr

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

11-04-2025 Data spaces: experience from the European Language Data Space
https://language-data-space.ec.europa.eu

Q&A

Flora Kopelou
data.europa.eu,
Publications Office of the EU

Philippe Gelin
Head of Sector
Multilingualism,
DG CNECT

Stelios Piperidis
Senior researcher, Head
of the Institute for
Language and Speech
Processing, Athena RC

data europa academy

Stay up-to-date on our **2025 activities!**

data.europa academy

Register now for our next webinar!

data europa academy

Continue the discussion on our collaboration channel!

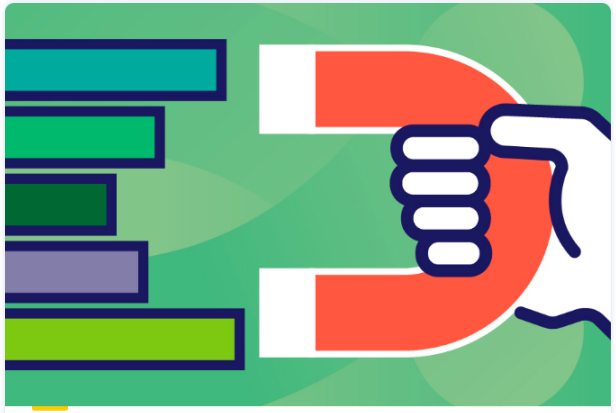
Collaboration channel

Connect with a vibrant community of data enthusiasts!

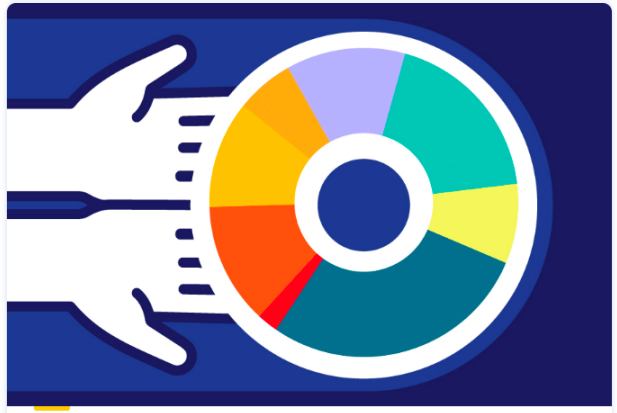
This space is designed for users to share ideas and exchange challenges and opportunities in the scope of the constantly evolving data landscape. You can join the group and topic you prefer, follow and be updated on ongoing conversations and participate in discussions on topics that matter to you.

Whether you are a data provider or a data reuser, you will find a dedicated space for collaboration where you can foster meaningful discussions.

If you are an official data provider, you can request access to this restricted forum, and we will carefully analyse your request. If you are a data reuser, you can join and will have direct access to this community to connect and engage with other members.



Data reusers



Data providers

data europa academy

Your opinion is important to us!

data.europa academy

Thank you!

data europa academy