

WEBINAR

Data spaces: experience from the European Language Data Space

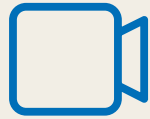
data.
europa
academy

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



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



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



The Language Data Space (and the ALT-EDIC)

Data.europa academy

Philippe GELIN
DG CONNECT – G3

50%
English

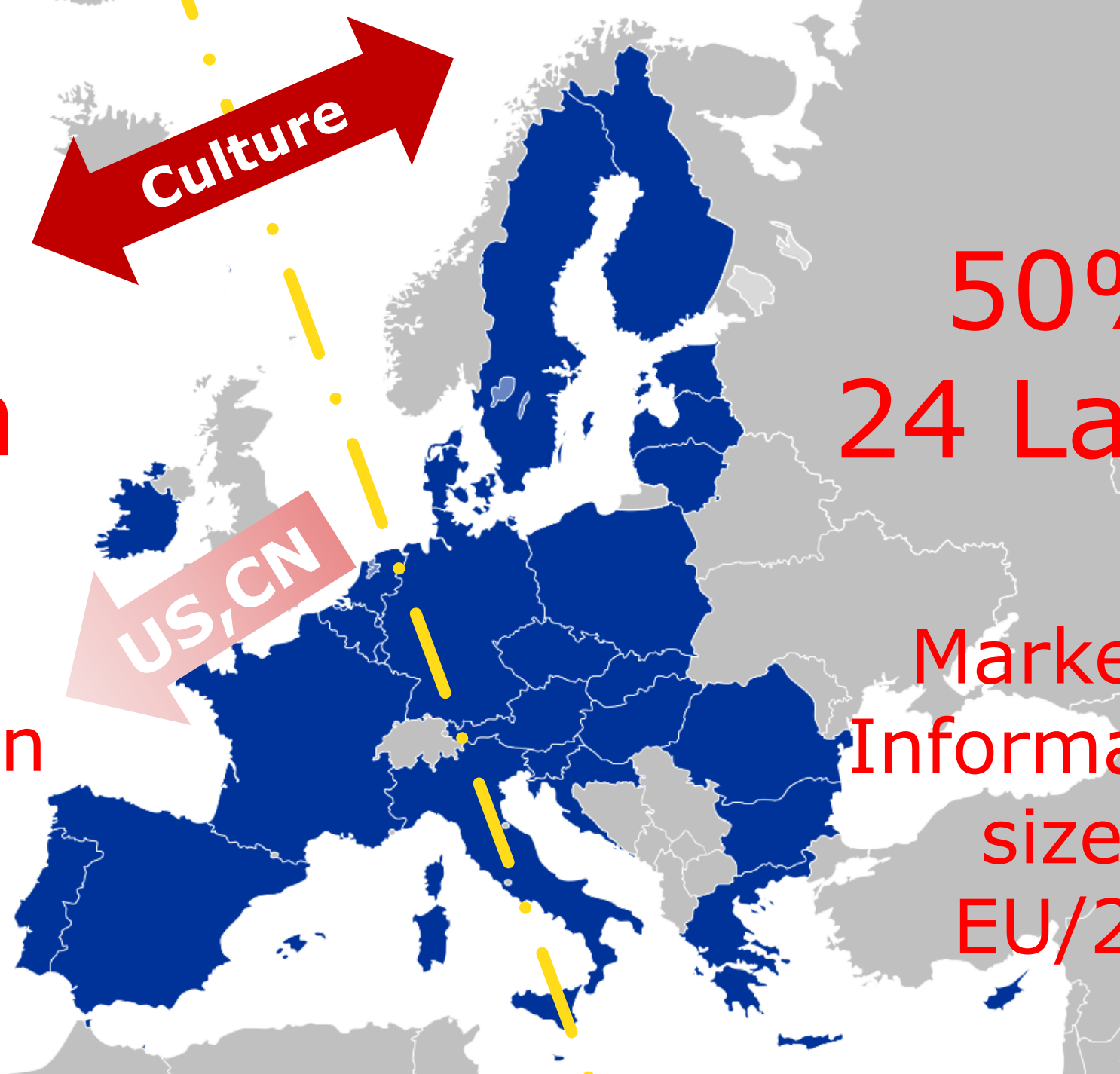


50%
24 Lang.

Market /
Information
size:
EU/2



Market /
Information
size:
EU/24

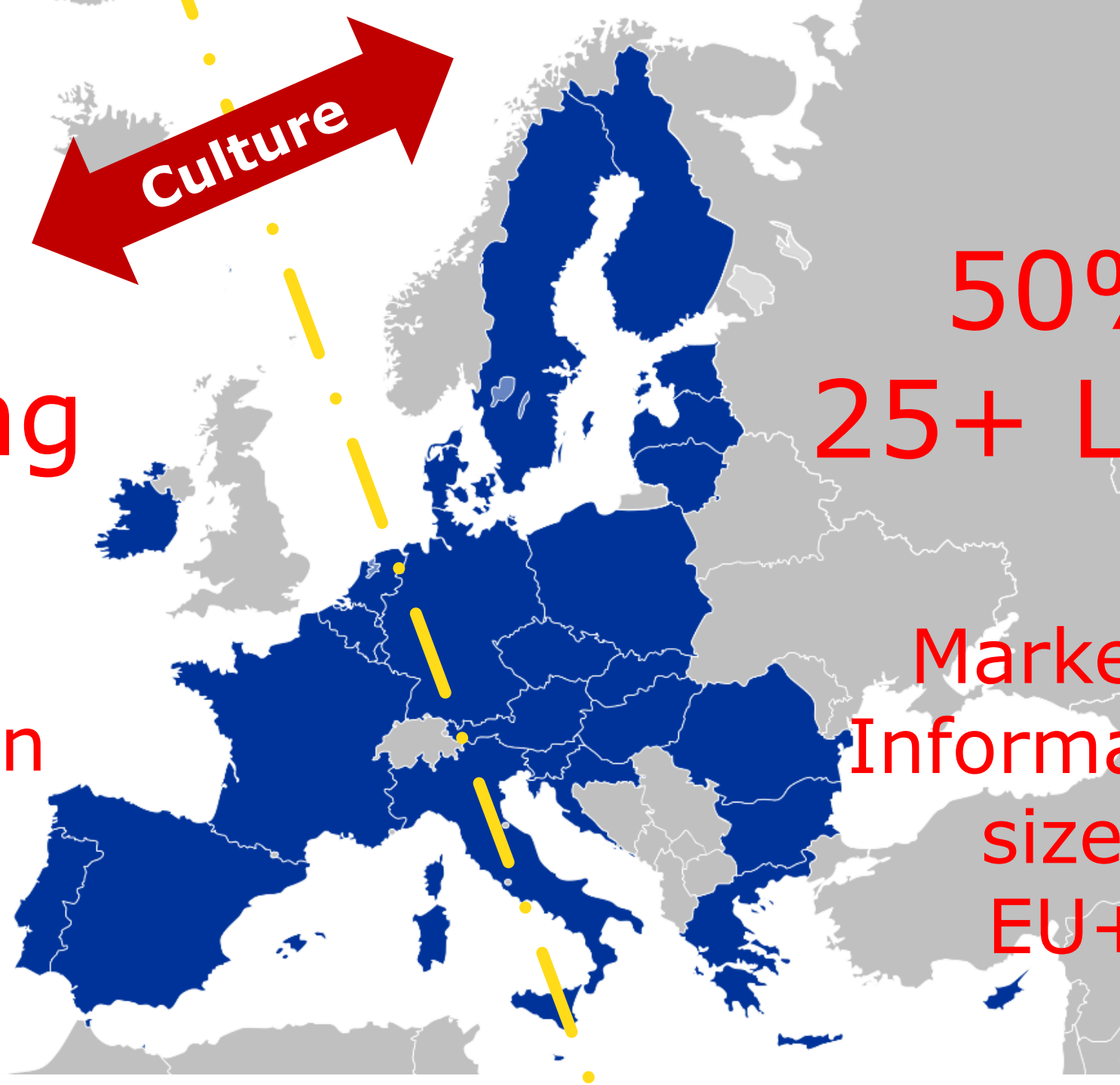


50%
25+ Lang

50%
25+ Lang

Market /
Information
size:
EU+

Market /
Information
size:
EU+





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.

Defining language technology

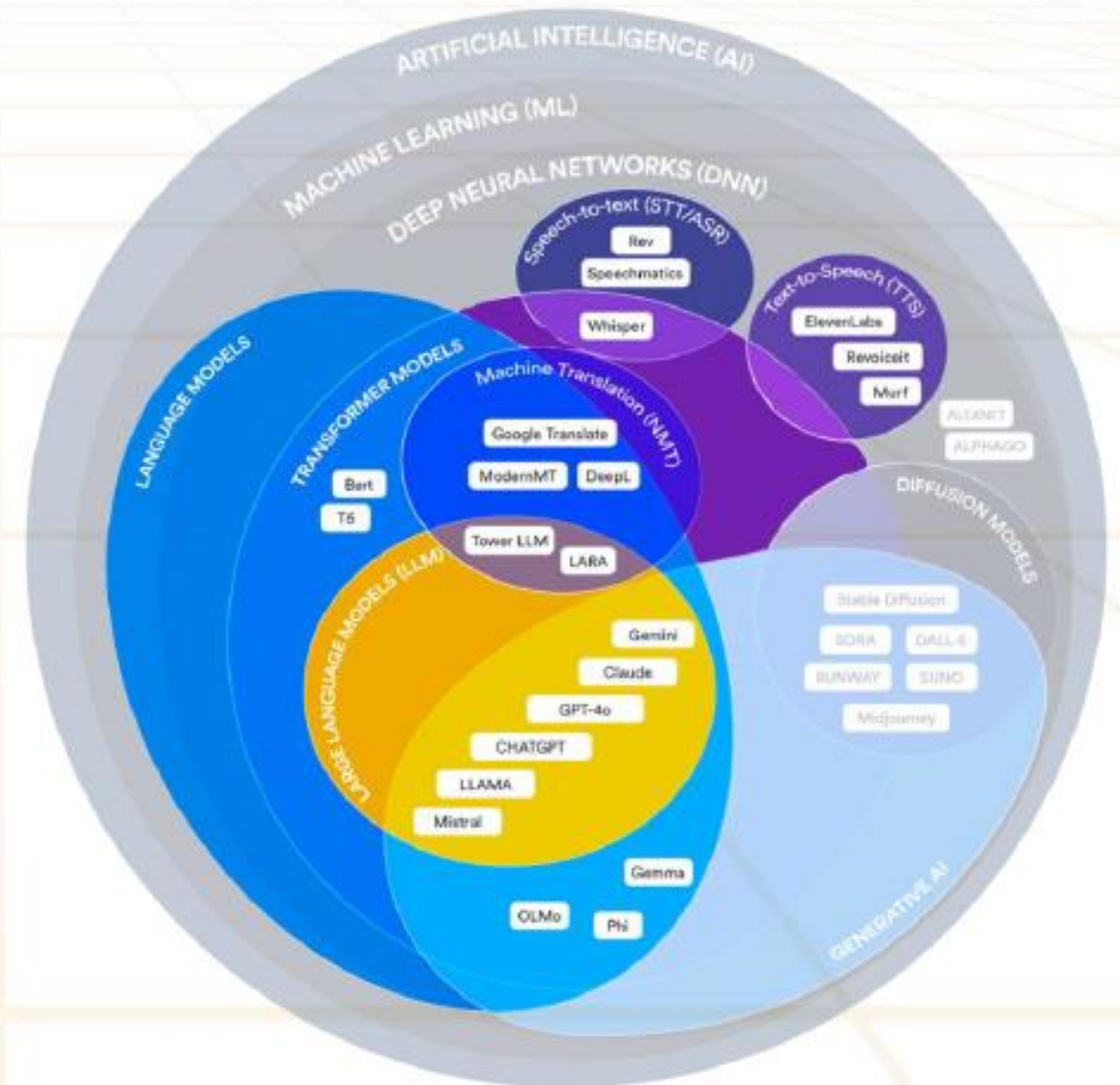
AI is eating the world of (language) technology

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)

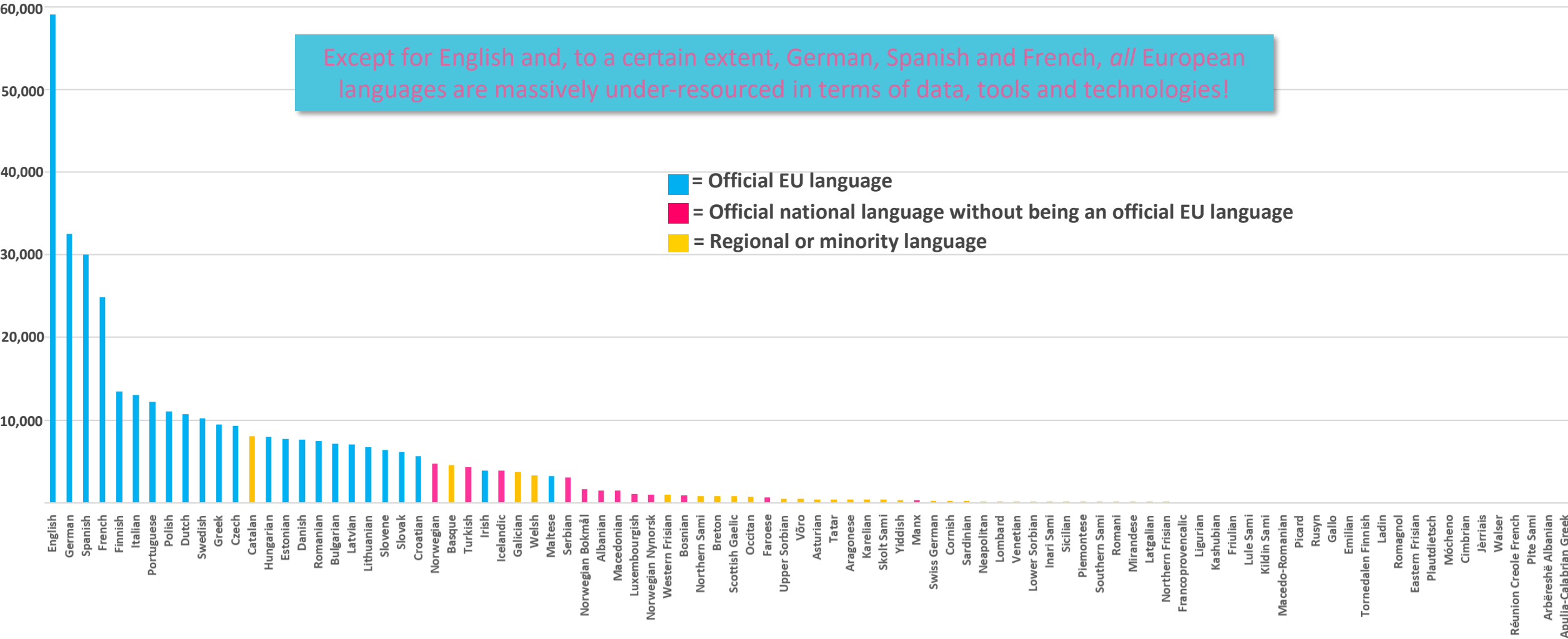
<https://www.nimdzi.com/the-language-ai-alphabet-transformers-llms-generative-ai-and-chatgpt/>

Digital Language Equality Metric: Technological Scores

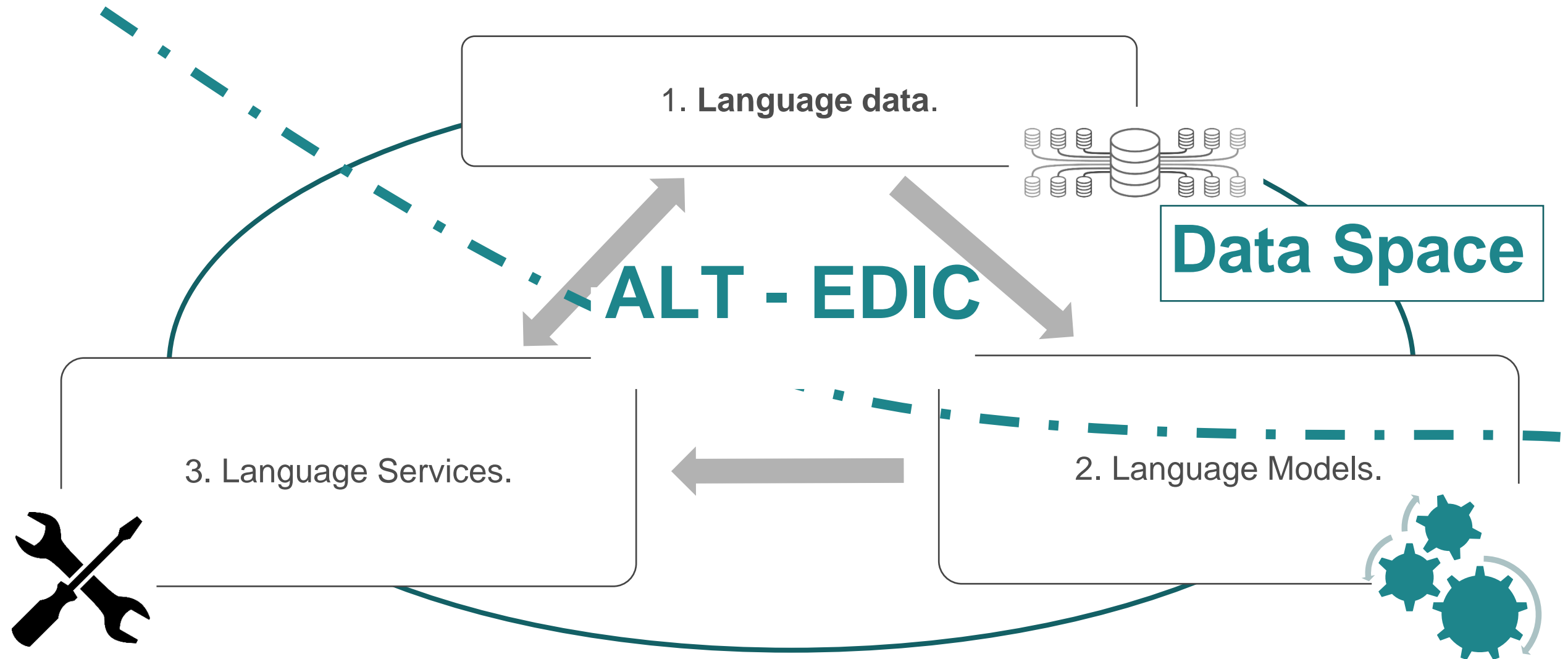


Except for English and, to a certain extent, German, Spanish and French, *all* European languages are massively under-resourced in terms of data, tools and technologies!

- = Official EU language
- = Official national language without being an official EU language
- = Regional or minority language



European Language Technologies Landscape



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

Federate MS and EU language data resources

Seed fund for new language models

Fine-tuning existing language models

Ease HPC access

Support SMEs take up

European Ecosystem

Budget

Projects (Costs)

- LLMs4EU	€ 40 Mio
- ALT-EDIC4EU	€ 4 Mio
- OpenEuroLLM	€ 40 Mio
- LLM-BRIDGE	€ 6,8 Mio

Common European Language Data Space



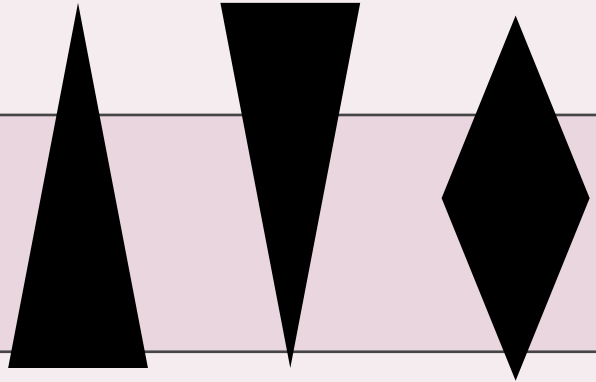
- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€ if renewed) – runtime: 36 months (+ 12 months if renewed)
- Objective: Develop and deploy a trustworthy and secure European infrastructure and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal 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



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	LLM 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;

Shaping Europe's digital future

[Home](#) | [Policies](#) | [Activities](#) | [News](#) | [Library](#) | [Funding](#) | [Calendar](#) | [Consultations](#) | [AI Office](#)[Home](#) > [Calendar](#) > Language Technology Landscape Conference 2025

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



© Nimdzi Insights

Organiser



EUROPEAN LANGUAGE DATA SPACE



The **European** Language Data Space Principles, Architecture and Components

Stelios Piperidis (Athena RC, Greece)

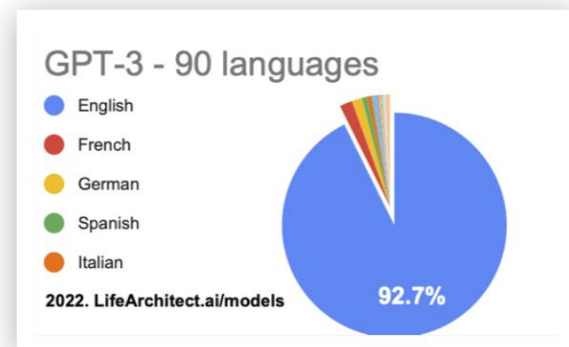
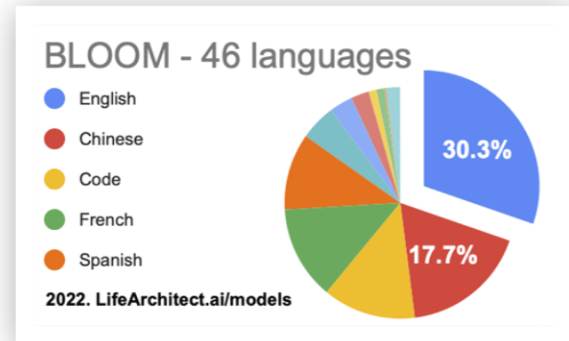
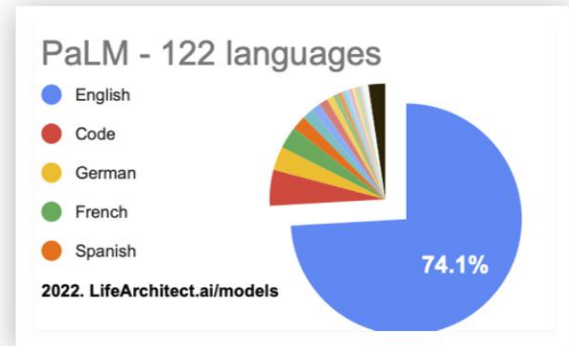
spip@athenarc.gr

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

META-SHARE LEARN • DISCOVER • PARTICIPATE • CONNECT • LOGIN

Search & exchange language resources

META-SHARE is an open and secure network of repositories for sharing and exchanging language data, tools and related web services

Share your own resources!

[JOIN OUR NETWORK NOW](#)

Already a member? [Log in](#)

Search the META-SHARE inventory

OR LEARN MORE

4,481 users | 2,887 language resources | 32% text corpora | 27,630 number of downloads

Virtual Language Observatory Search Contributors Help CLARIN

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or **continue** to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

[See all records](#) [Take a quick tour](#)

Search through 1,030,321 records

European Language Resource Coordination Connecting Europe Faculty ELRC-Share

ELRC-SHARE Repository

Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources Coordination and considered useful for feeding the CEF Automated Translation (CEF AT) platform.

If you want to contribute resources, all you have to do is [register](#) (new user) or [login](#) (returning user) and go on to

European Language Grid

Language Technologies

Discover, try out, use and download LT services and resources for all European languages.

Browse ELG and find the LT services, resources, developers and providers you are looking for.

Search the catalogue

8000 Corpora | 3884 Tools & Services | 2812 Conceptual Resources | 510 Models & Grammars | 1775 Organizations | 513 Projects



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



- 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

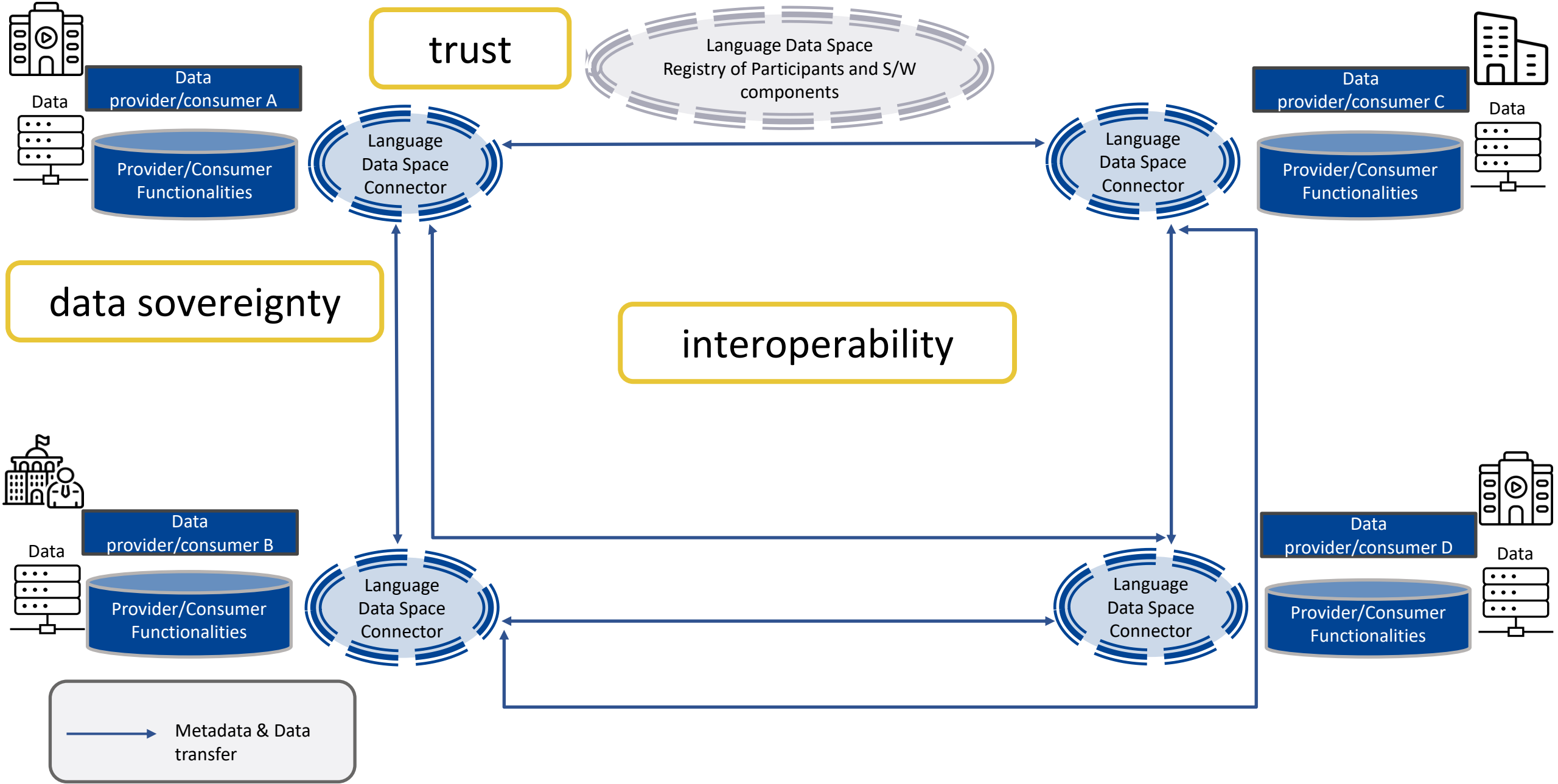
data sovereignty

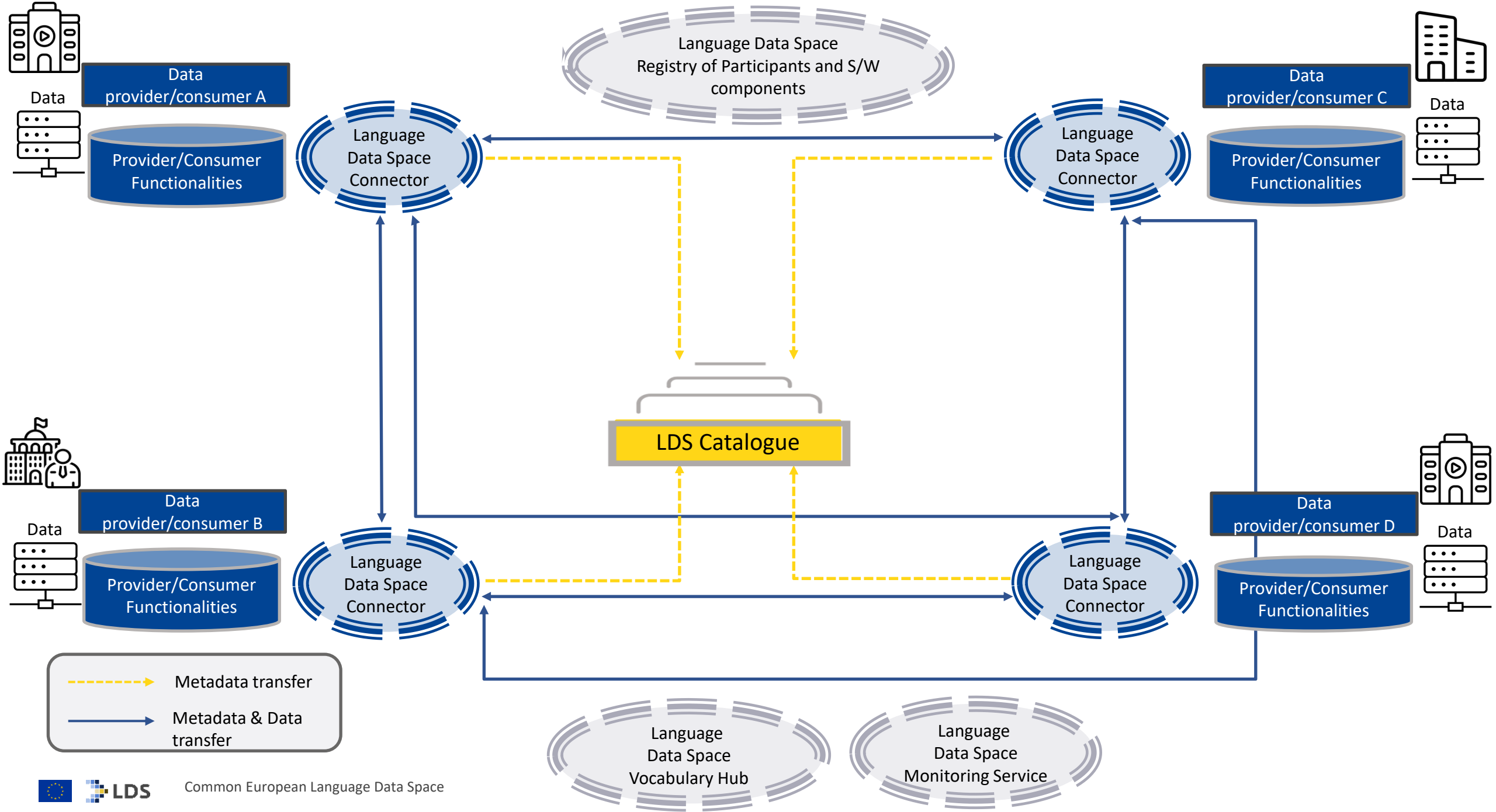
trust

interoperability

LDS Architecture

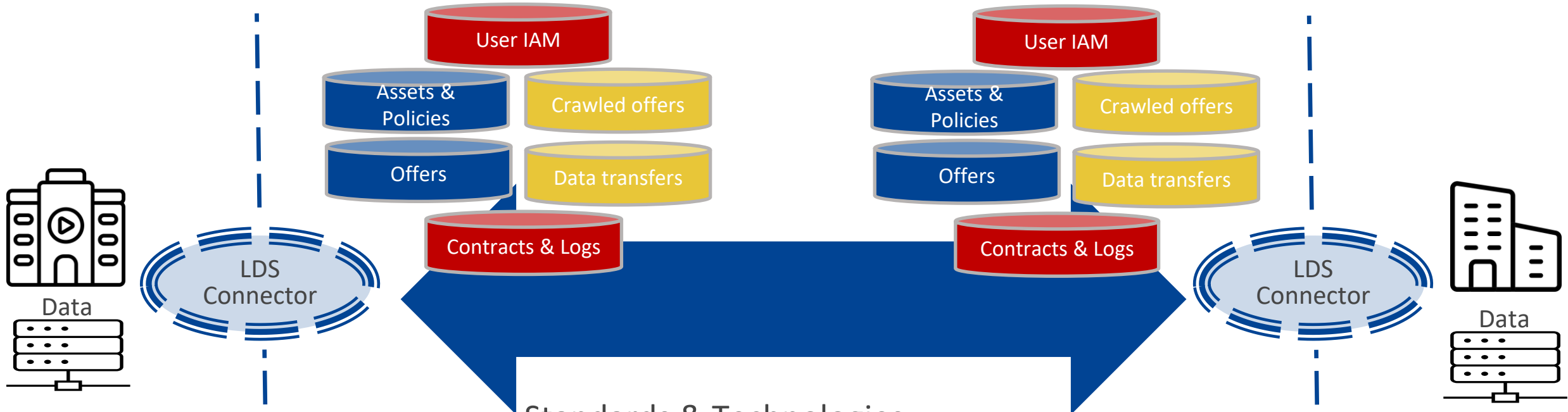
- Objective 1: **provide technological support for complying with the Data Spaces principles**
 - sovereignty, trust, interoperability
- Objective 2: **provide technological support for LDS workflows** designed
 - for data providers & consumers
 - for the LDS Governance Board





LDS Connector

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



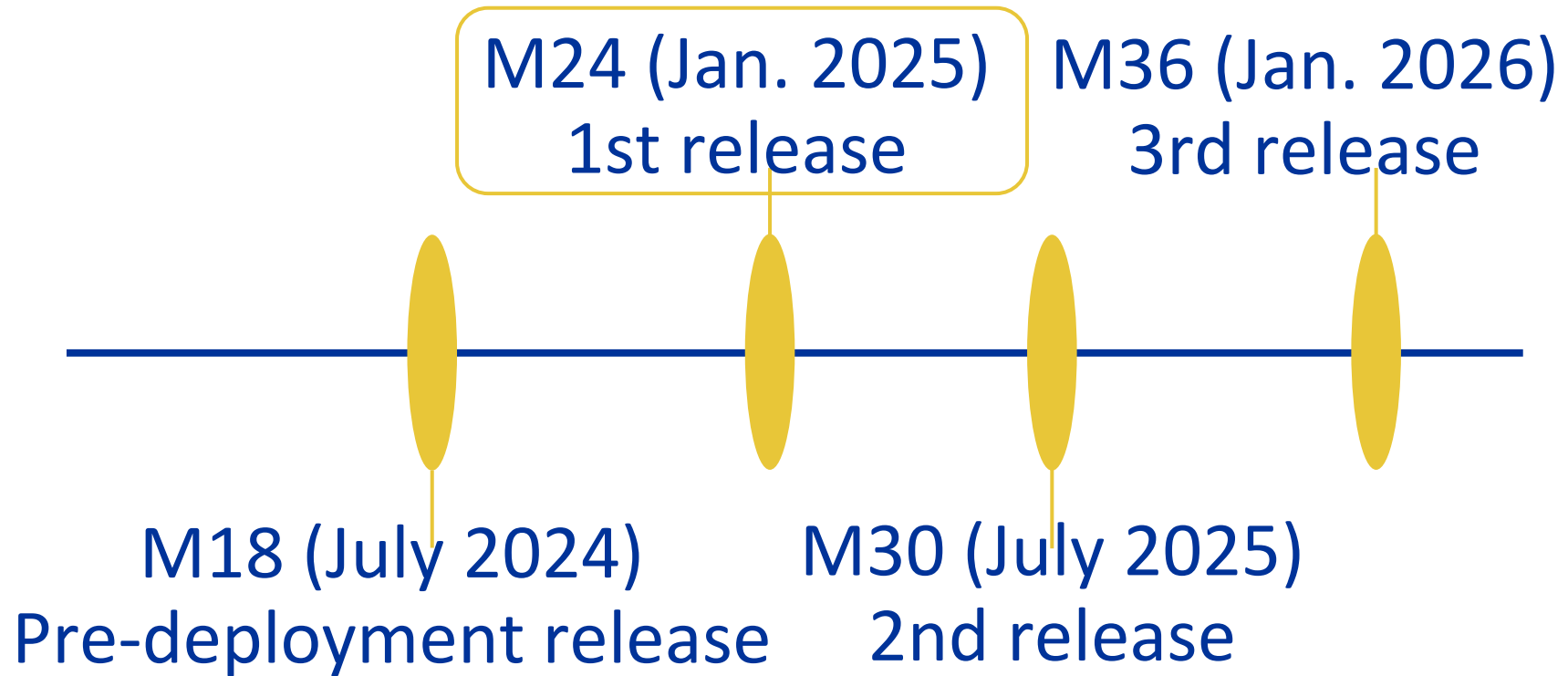
Standards & Technologies

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

Connector features for

- data providers
- data consumers
- data prosumers

LDS Release Plan





**EUROPEAN
LANGUAGE
DATA SPACE**

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**

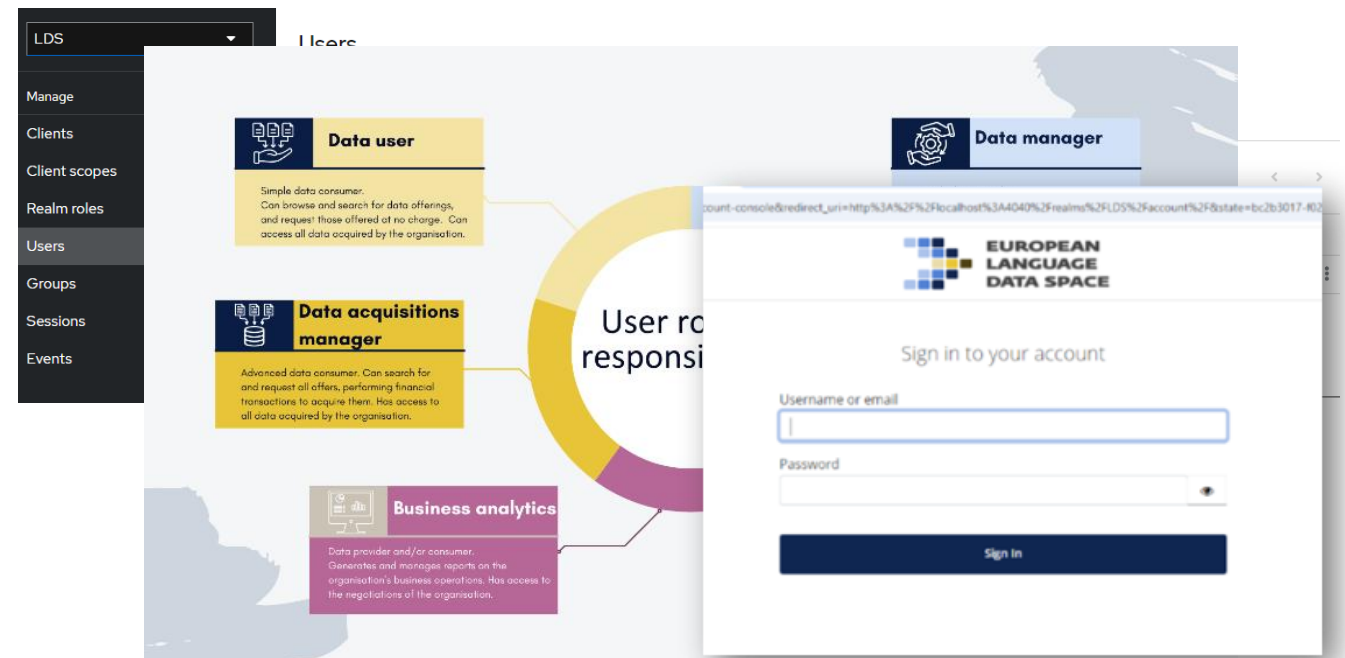
[Demo 1: Apply for a membership in the LDS - YouTube](#)



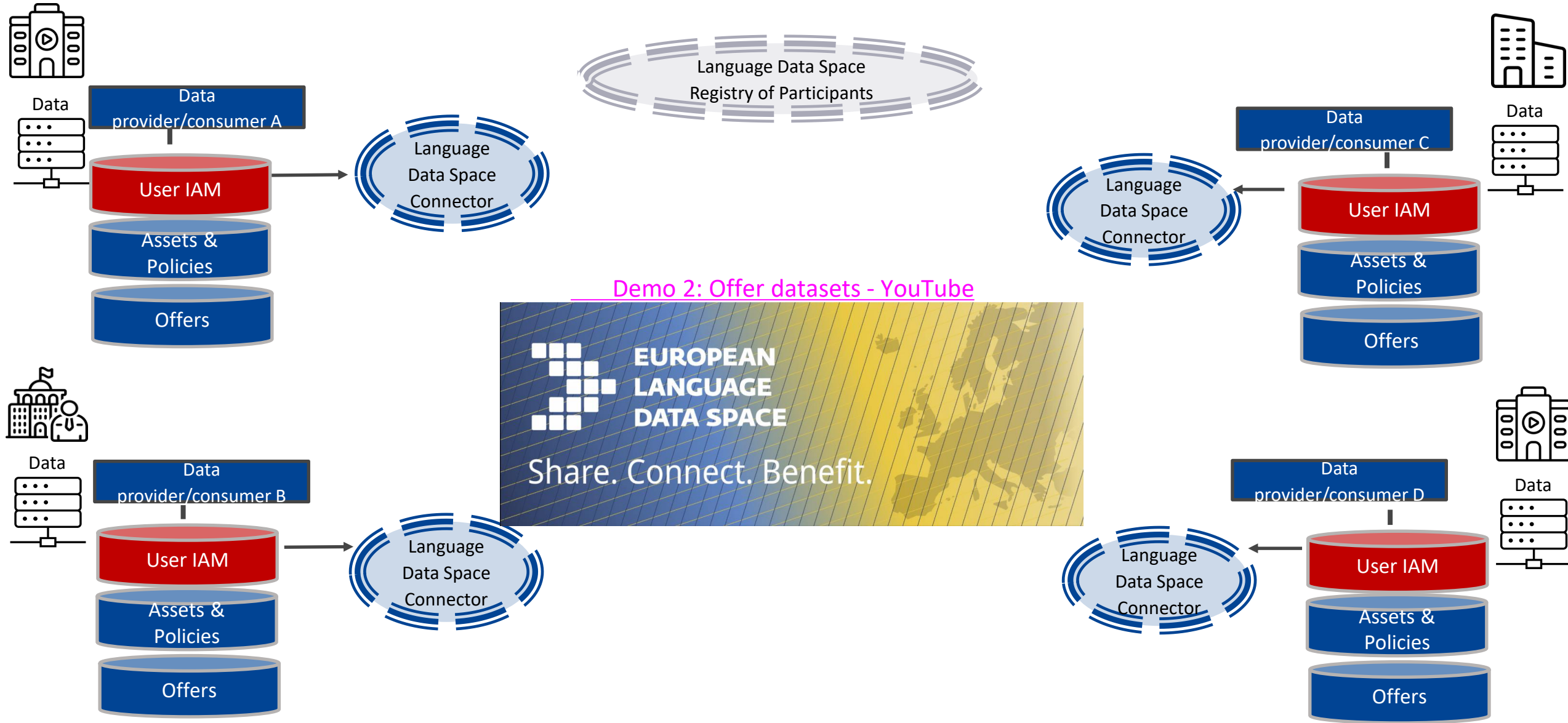
2. User Identity Access Management (inside the participant's LDS connector)



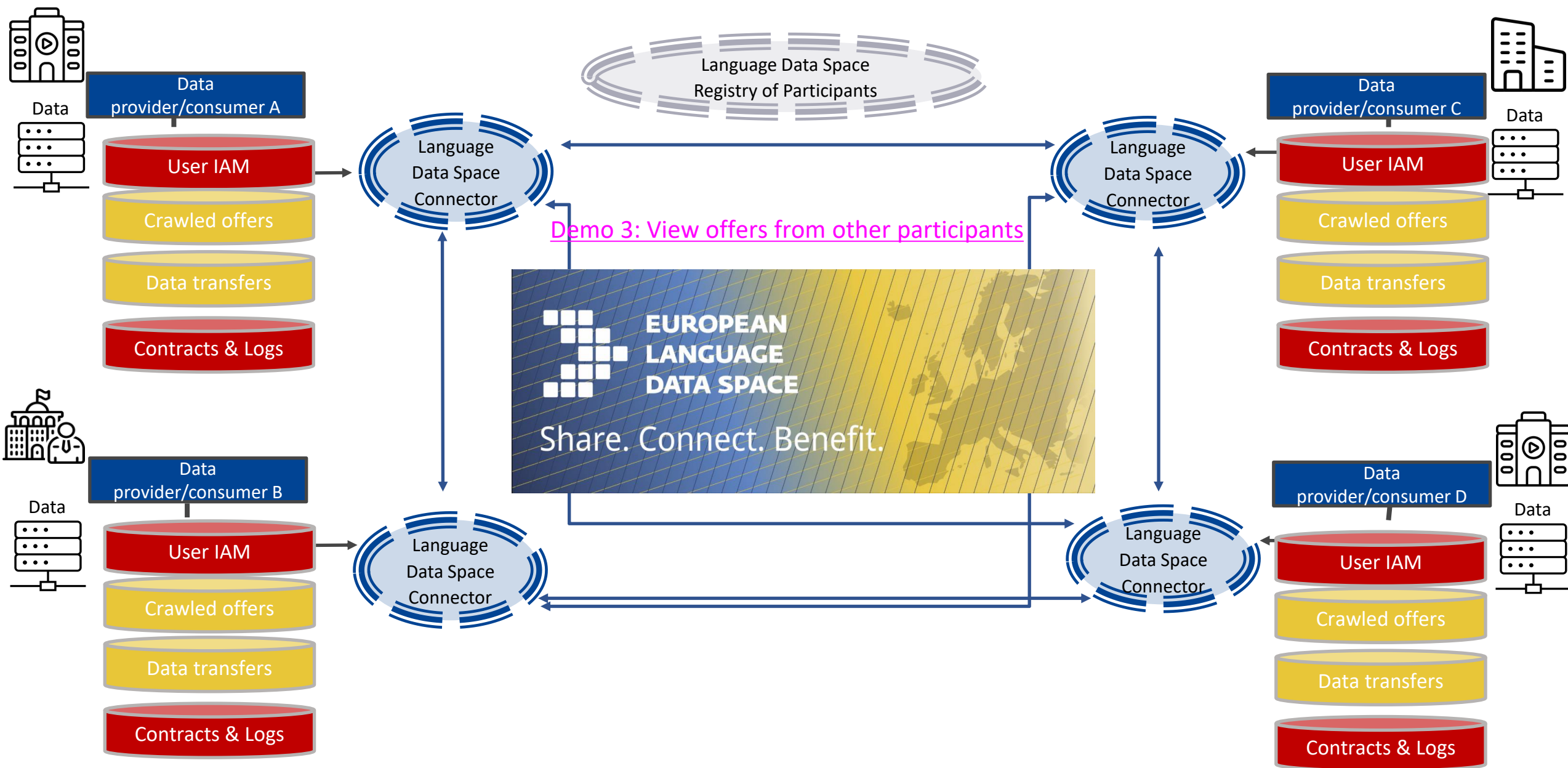
- **Keycloak** solution
- **Registration** of users by user administrator (with possibility to use active directory)
- Assignment of pre-defined LDS **user roles** for providers & consumers with simple and advanced rights (with possibility to assign multiple roles)



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

Help



© Freepik

NAVIGATION
Contact form
Frequently Asked Questions

For any technical issue regarding the website, for legal or technical questions concerning the Language Data Space, or if you have any questions or concerns, please don't hesitate to reach out to us by using the contact form below.

A number of pre-set topics is available. Summarising your request in the subject field will help us involve the staff best suited to address your needs.

Before contacting us, you might want to explore our [FAQ section](#) to access further information and documentation.

Home > Help > FAQ

FAQ

LDS documentation User Guide

GitHub ⚙️



Share

You will be able to share and monetise your language data, language models and other language resources through a single platform, taking European values and compliance with EU regulations fully into account.



Connect

You will be able to connect and exchange with other stakeholders through the European Language Data Space.



Benefit

You will be supported by the European Language Data Space in the development of multilingual and multimodal language technologies and language-centric AI.

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

The screenshots show the 'LDS Connector Management Panel' interface. The top-left panel displays 50 Assets, 18 Policies, and 25 Offers. The top-right panel displays 60 Assets, 21 Policies, and 29 Offers. The bottom-right panel displays 6 Assets, 22 Policies, and 10 Offers. A central yellow diamond shape is formed by three yellow trapezoidal shapes pointing towards it. The bottom-right screenshot also includes a 'Language Technologies' section with a search bar and a table of statistics.

Category	Count
Assets	8000
Books & Journals	3884
Language Resources	2812
Metadata Sources	510
Repositories	1775
Projects	513

Apply at

<https://ldssetup.ilsp.gr/> to

join LDS and test it

vanilla

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>



Join the LDS user group

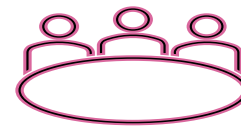


© Freepik

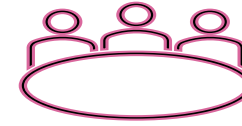
The European Language Data Space (LDS) user group members shall actively contribute to and take advantage of the LDS, bringing in their own requirements and validating the emerging LDS infrastructure.

If you are a stakeholder who is in need of language data or if you want to give the language data of your organisation a second life, potentially monetising it, you are welcome to join.

Click to join



LDS Connector



Metadata & Data



Policies



Common European Language Data Space

Thank you!



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

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>

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



Stay up-to-date on our
2025 activities!

The logo for Data Europa Academy is located in the bottom left corner. It consists of the words "data", "europa", and "academy" stacked vertically in a white, lowercase, sans-serif font. The word "data" has a small yellow dot above the 'a', and "europa" has a small yellow dot above the 'o'. The word "academy" is in a smaller font size. The logo is set against a dark blue circular background, which is part of a larger graphic design featuring overlapping purple and blue circles.

data.
europa
academy

Register now for our next webinar!

WEBINAR

Unlocking data interoperability:
exploring ways of
standardisation
with real use cases

data.
europa
academy

25 April 2025

10.00 – 11.30 CEST



data.
europa
academy



Continue the discussion on our collaboration channel!

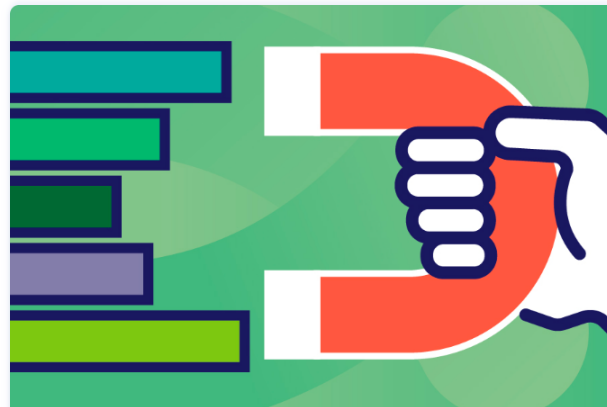
Collaboration channel

Connect with a vibrant community of data enthusiasts!

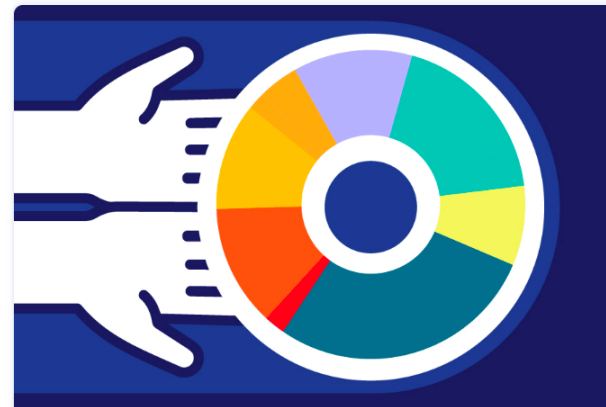
This space is designed for users to share ideas and exchange challenges and opportunities in the scope of the constantly evolving data landscape. You can join the group and topic you prefer, follow and be updated on ongoing conversations and participate in discussions on topics that matter to you.

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.



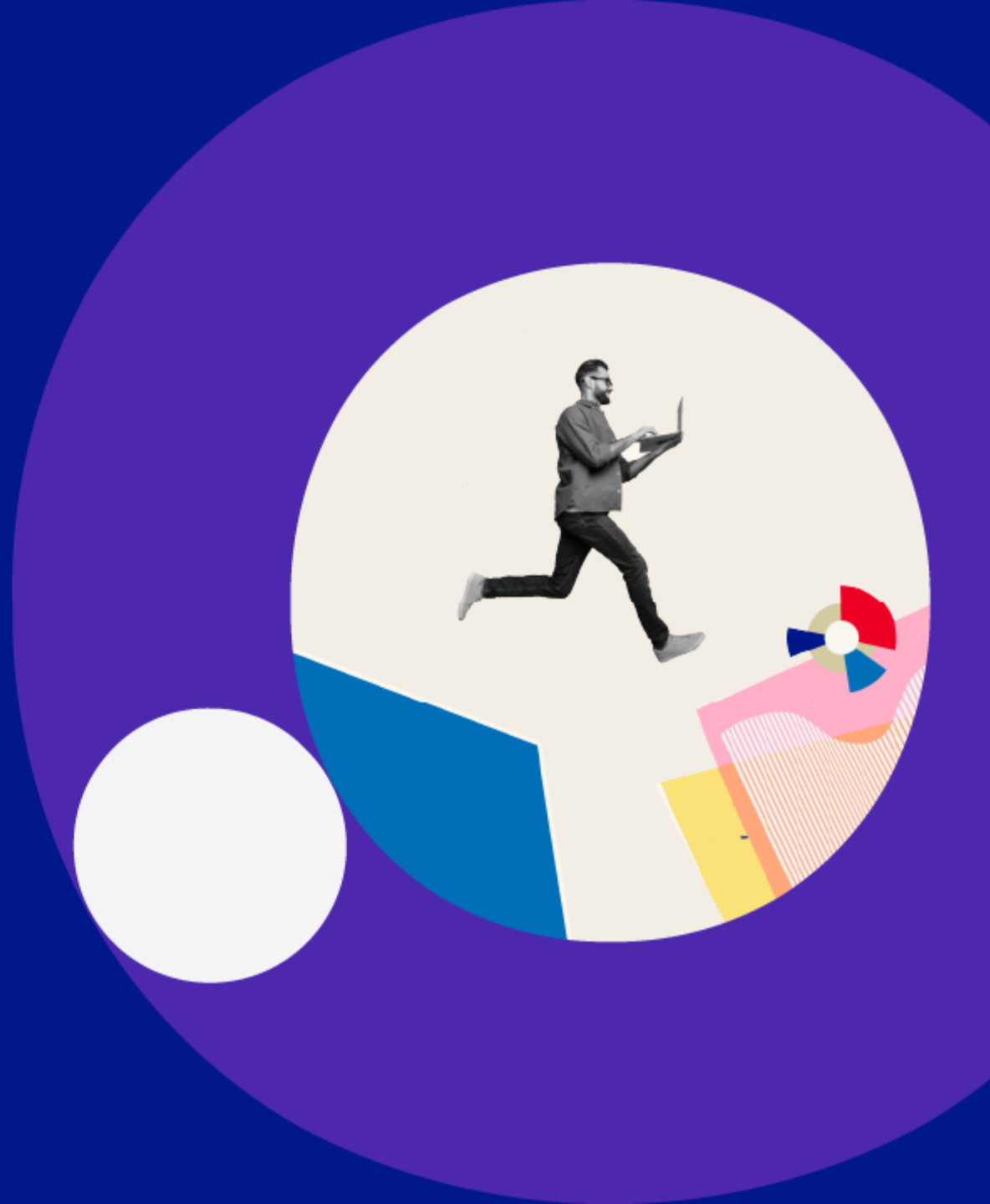
Data reusers



Data providers



Your opinion is important to us!



Thank you!

