



# EUROPEAN LANGUAGE DATA SPACE



## LDS Country Workshop Belgium & The Netherlands

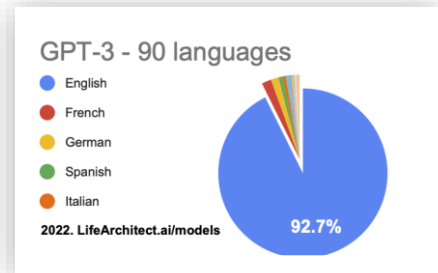
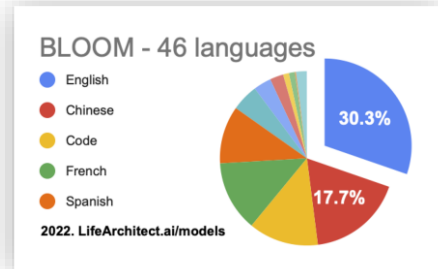
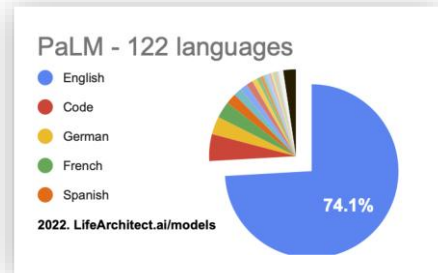
Katrin Marheinecke, Georg Rehm (DFKI GmbH, Germany) – LDS Project Coordination  
katrin.marheinecke@dfki.de

23-01-2025 LDS Country Workshop Belgium and Netherlands

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

# Context: Large Language Models (LLMs)

- Unprecedented capabilities: LLMs are the **most disruptive breakthrough** in AI in recent history (GPT-3, ChatGPT, GPT-4, Claude, Gemini etc.)
- Trained on **vast amounts of data** (+ images, videos, audio, i.e., multimodal data)
- LLMs are getting larger and larger: more *data*, more *parameters*, more *compute*. - > **Scaling laws**: larger models outperform smaller models.
- 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**
- Already now billions and billions are made but Europe lags far behind

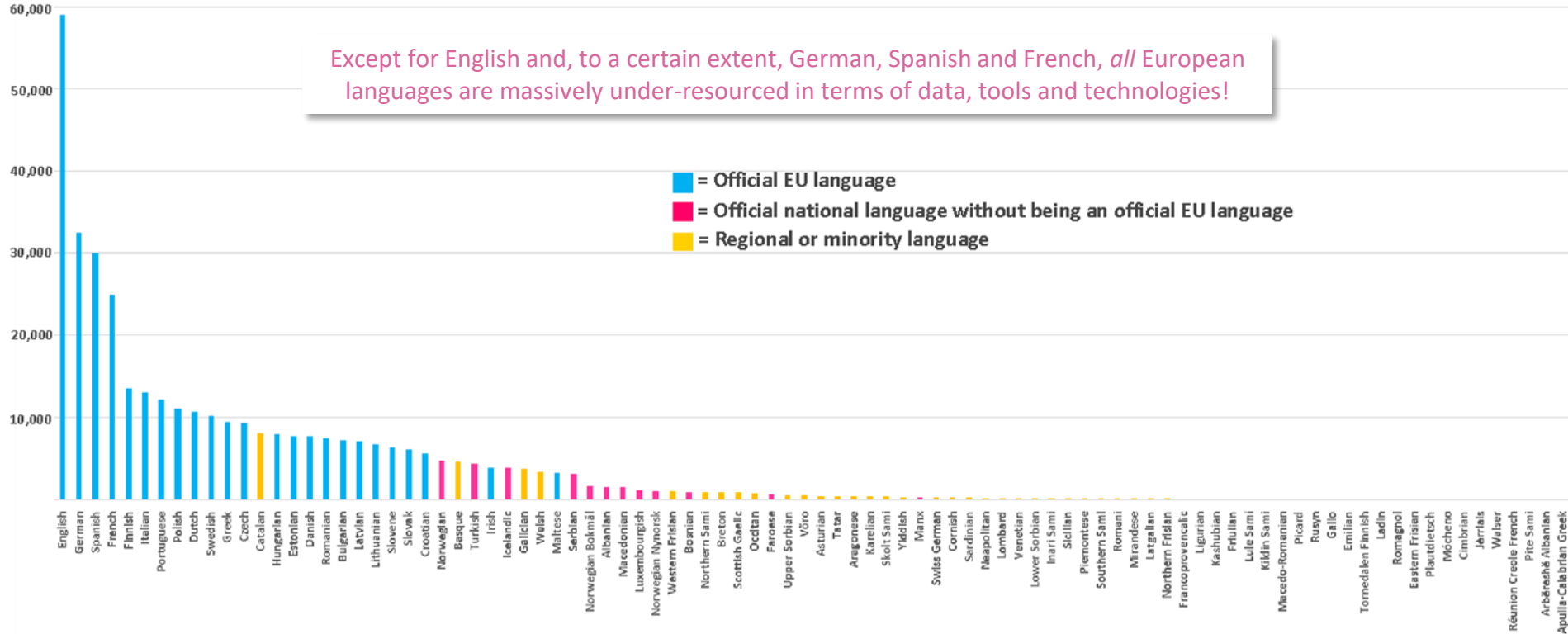


# European Initiatives

- 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 and nationally funded projects, e.g., HPLT, TrustLLM
  - New pan-European initiative: ALT-EDIC
- Challenges:
  - Availability of data for European languages
  - High Performance Computing facilities (HPC)
  - Speed of the big tech players in the US and Asia vs. speed of Europe

# 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!



# EU Data Strategy

# 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

## 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** that are **Coordinated** 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** (existing user) and go on to describe

European Language Resource Coordination

## ELRA REPOSITORIES

1096 Language Resources (Page 1 of 55)

2006 CoNLL Shared Task - Arabic & Czech

Order by: Resource Name A-Z

Year	Language	Resource Name	License
2006	Arabic	Arabic News	CC-BY
2006	Czech	Czech News	CC-BY
2006	Arabic	Arabic News	CC-BY
2006	Czech	Czech News	CC-BY

**Ten Languages**

Arabic, Japanese, Portuguese, Slovenian, Spanish, Swedish, Turkish

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.

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

# EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy
- Data Spaces will help to establish a data economy in Europe
- Various data economy and data infrastructure initiatives in Europe with slightly different goals and individual positioning but conceptual, technical, legal and operational overlap:
  - Data Spaces Business Alliance (DSBA): Gaia-X, IDSA, FIWARE, BDVA
  - EU: DSSC (incl. DSBA), Simpl, approx. 20 data spaces
- The **Common European Language Data Space** is one of the official EU data space projects with a strong focus on industry

# Common European Language Data Space

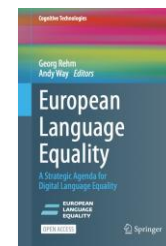
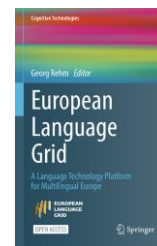


- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€ if renewed)
- Runtime: 36 months (+ 12 months if renewed)
- Objective: Develop and deploy a European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- Salient features: governance framework, technical infrastructure, openness, promotion of all languages
- Stakeholders: **industry**, research, public administration, cultural associations, NGOs and citizens
- The four core partners have been involved in many projects, including:

- META-NET (FP7, 2010-2013): META-SHARE
- ELRC (CEF, 2014-2023): ELRC-SHARE
- ELG (H2020, 2019-2022): ELG Cloud Platform
- ELE (PP/PA, 2021-2023)



The **technical development work in LDS** will be informed by ELG, ELRC-SHARE, META-SHARE.





# Consortium and Subcontractors

<b>Lead Partner and Coordinator</b>		
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	DE
<b>Partners and Operation Leads</b>		
R.C. "Athena", Institute for Language and Speech Processing	ILSP	GR
Evaluations and Language Resources Distribution Agency	ELDA	FR
TILDE	TILDE	LV
<b>Main Subcontractors</b>		
3pc GmbH Neue Kommunikation	3pc	DE
CLARIN ERIC	CLARIN	NL
Big Data Value Association (Data, AI and Robotics) AISBL	BDVA	BE

Plus legal experts (Delcade, France) and approx. 30 organisations for the logistics of multiple country workshops

# Classes of Data

Class of Data	Typical Size	Providers	Integration into LDS	Relevance for LLMs
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 other sectors that own language data	Can be easily integrated by connecting these organisations to LDS	Especially high quality data or domain-specific data or data covering specific languages and thus highly relevant for LLMs

# Alliance for Language Technologies EDIC (ALT-EDIC)

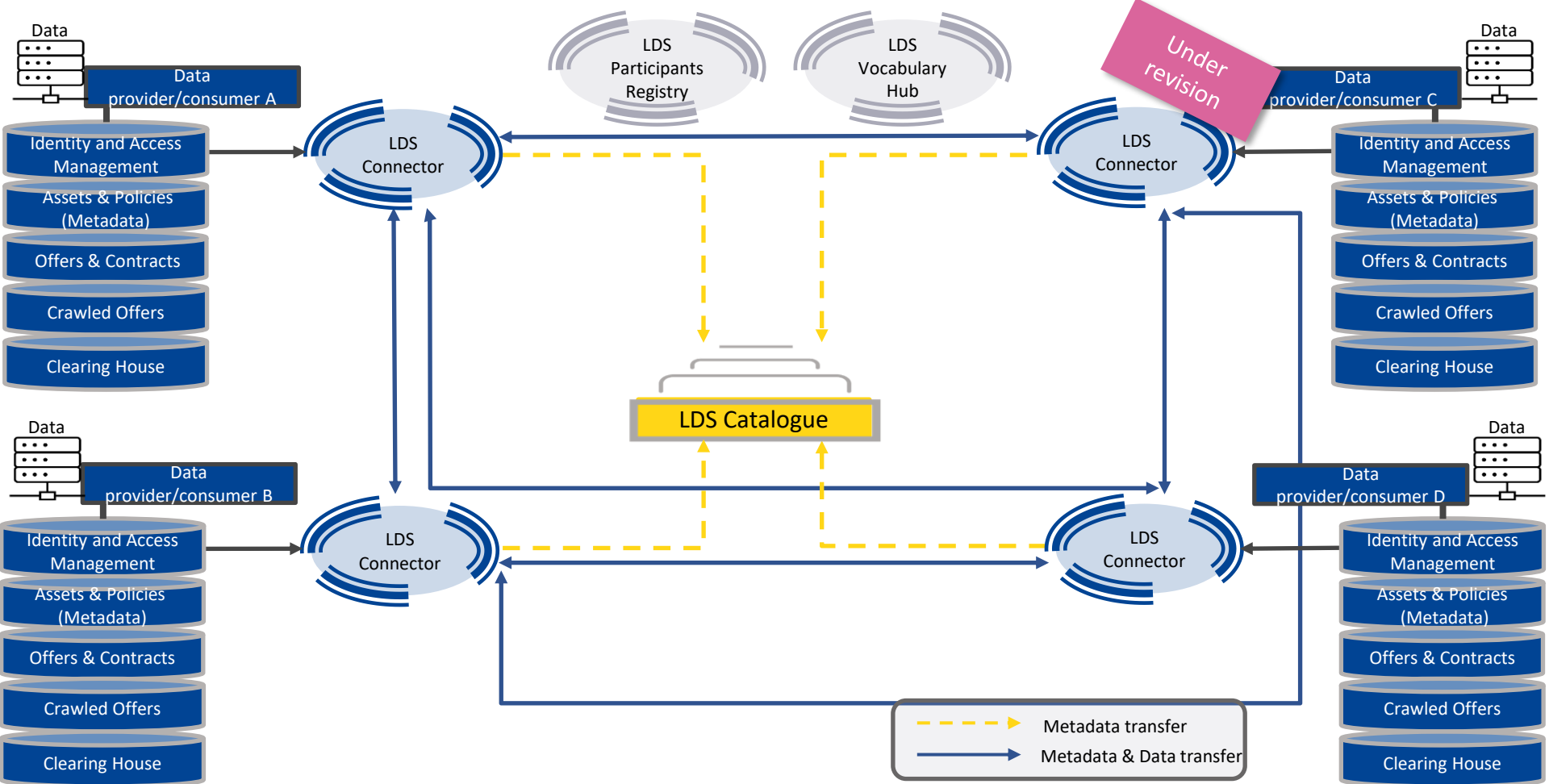
- European Digital Infrastructure Consortium (EDIC): a new legal entity type in the EU
- The first couple of EDICs are currently under development including the ALT-EDIC
- Coordinated by the French Ministry of Culture
- Close collaboration between: ALT-EDIC Working Group, EC, LDS
- ALT-EDIC action plan will concentrate on:
  - 1. Data;
  - 2. Existing language models;
  - 3. New language models;
  - 4. Evaluation, certification, normalization;
  - 5. Ecosystem;
  - 6. EDIC implementation
- We expect many synergies between LDS, ALT-EDIC, DSSC, Simpl, other data spaces and other projects!

## ALT-EDIC Members

**17 Members States:** Bulgaria, Croatia, Czechia, Denmark, Finland, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Slovenia and Spain;

**8 observing Member States:** Austria, Belgium, Cyprus, Estonia, Malta, Portugal, Romania and Slovakia.

# The Common European Language Data Space - Implementation



Early LDS  
Prototype



MANAGEMENT

[Home](#)

History ▼

Storage Solutions ▼

OPERATIONS

Assets ▼

Policies ▼

Offers ▼

## LDS Connector Management Panel

Here you can create and manage your assets, your policies and your offers and review your contract agreements.

<p>Assets <b>15</b></p> <p><a href="#">Create A New Asset</a> <a href="#">View My Assets</a></p>	<p>Policies <b>16</b></p> <p><a href="#">Define A New Policy</a> <a href="#">View My Policies</a></p>	<p>Offers <b>6</b></p> <p><a href="#">Create Offers</a> <a href="#">View Offers</a></p>
------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------

# Step 1: Create new data asset

Create a new asset

Select language (optional)  
Language  
ENGLISH

**Basic properties**  
Title, short description, version, ...

**Privacy properties**  
anonymization or sensitive data details.

**Language**  
A language of the resource.

**Type properties**  
media type, linguality type, annotation type, corpus subclass, ...

**IprHolder properties**  
iprHolder or creator details...

**Documentation**  
related documentation

**Temporal properties**  
time constraints

**Identifiers**  
identifier details

**Distribution**  
media type, format, ...

**Data address**  
base url, type, ...

Details

Privacy

Language

Type

IPR holder

is documented by

Temporal Coverage

Identifiers

distribution

Data address

SAVE ASSET

Early LDS  
Prototype

## Step 2: Create and adjust policy

### POLICY CLASS

Early LDS  
Prototype

<p>✓ Interval-restricted Data Usage Policy Class allows data usage for a specified time period</p> <p>pending</p>	<p>Purpose-restricted Data Usage Policy Class allows data usage for a specific declared purpose</p> <p>pending</p>
<p>Connector-restricted Data Usage Policy Class allows data usage for a specific connector</p> <p>pending</p>	<p>Perpetual Data Sale (Payment once) Policy Class allows data usage after payment is completed (for datasets requiring once-off payment)</p> <p>pending</p>
<p>Location Restriction of the participant for Data Usage Policy Class restricts data usage to participants in a specific location</p> <p>pending</p>	<p>Attribution Data Policy Class allows distribution of data with mandatory attribution</p> <p>pending</p>
<p>Share Alike derivatives Policy Class allows distribution of derivatives with a compatible</p> <p>pending</p>	<p>Attach Policy When Distribute to a third-party Policy Class allows distribution of data to a third-party with the specified attached policy</p> <p>pending</p>
<p>Derivatives not allowed Policy Class distribution of derivatives is not allowed</p> <p>pending</p>	



## Step 3: Create and publish offer

Early LDS  
Prototype

Progress: ✓ Select Asset — ✓ Assign Policy — 3 Review & Publish

[Previous](#) [Publish](#)

Name	Description	LR type
Arab-Andalusian music corpus	This repository contains Arab-Andalusian corpus collected in the CompMusic project. The following files are available for 164 concert recordings (overall playabl...	Corpus
AcCompl-it Dataset	The AcCompl-It dataset comprehends the Complexity and the Acceptability Datasets. The first data set is composed of 2,530 Italian sentences annotated with human...	Corpus

Apache License, Version 2.0 ▾

# The Common European Language Data Space – a Value Proposition

# Language Data – Language Resources – Data Products

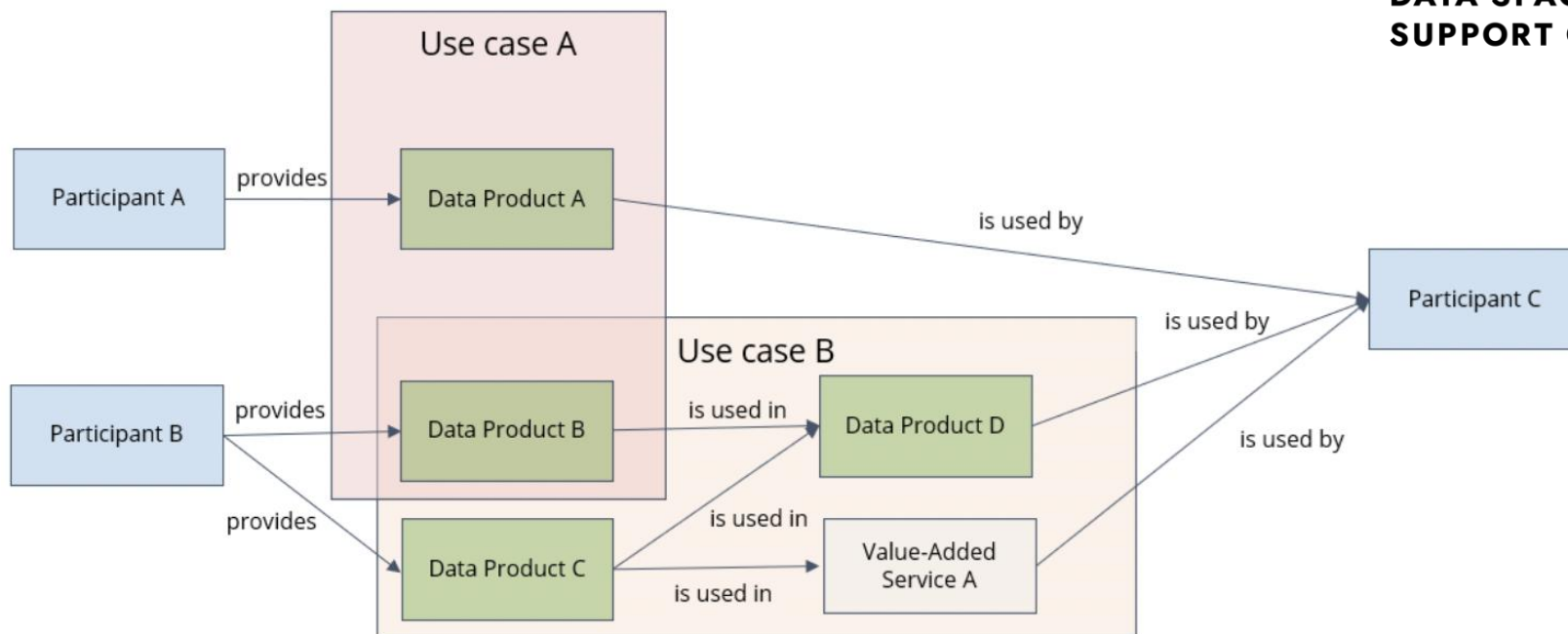
- The NLP and Computational Linguistics community has been sharing language data since the 1990s
- Back then: annotated corpora, treebanks, grammars, lexicons, smaller language models
- The term “language resource” (LR) was established (data, documentation, evaluation, metrics etc.)
- *language resource*  $\approx$  *data product*
- Of utmost importance now: In the LDS, identify and make available **large amounts of language data** to enable industry and research **to pre-train large language models for Europe**
- Typical availability of LRs *since the late 90s and early 2000s*:
  - For research purposes: free of charge
  - For commercial use: often for a certain fee. LRs developed by European research organisations were licensed by European but also large US tech companies, e.g., for online NLP services (Machine Translation). **LDS: Streamline exchange of such data**



# LDS and Data Products



**DATA SPACES  
SUPPORT CENTRE**



## DSSC Blueprint 1.0 – Data Product Development

# Language Data Space – Value Proposition

## Data Providers

- Additional revenue – LDS as a marketplace
  - Sell data products
  - 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 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



TAP RUNNETH DRY | 11.13.23, 4:05 PM EST by MAGGIE HARRISON DUPRÉ

## AI Companies Are Running Out of Training Data

The well is running dry.

[/ Artificial Intelligence](#) / [/ AI](#) / [/ AI Industry](#) / [/ AI Training](#)



Big Tech

# X won't train Grok on EU users' public posts

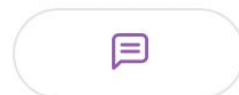
The permanent agreement follows pressure from data protection officials.



**Kris Holt**

**Contributing Reporter**

Updated Wed, Sep 4, 2024 · 1 min read



The company X.AI will stop training its model Grok on EU users' public posts – this will be a further problem for digital language equality because these posts may be written in (very) under-resourced languages. Maybe a trend?

# Data Products in LDS – Training Data for Generative AI and LLMs

- A few examples of recent agreements:
  - Reddit: \$60 million per year (Google)
  - Shutterstock: \$25-50 million (Apple)
  - Springer: Tens of millions (Open AI)
  - Offer for news publishers: \$1-5 million per year (Open AI)
  - Offer for owners of large datasets: \$50 million (Apple)
- Global market is **enormous** – owners/providers of large amounts of content **are paid large sums by the US technology enterprises** that currently dominate the AI product landscape
- January 2025: US Government announces \$500 billion of funding for **new AI initiative Stargate**
- It's up to the data providers to establish offers and prices that make sense for them
- Our ambition: to establish LDS as a **marketplace for European language data**



# LDS User Group

## • Meetings and Conferences

- Inaugural meeting in March 2024
- Second meeting in June 2024
- Third Meeting November 2024
- LDS Launch Conference: 19 March 2025

## • Communication

- Established a mailing list for the LDS user group
- The LDS user group will grow – new members are being added to the mailing list regularly
- **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; first trial exchanges of data; surveys etc.
  - You can also help on a more substantial, in-depth level – please approach us if you're interested.

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](#)



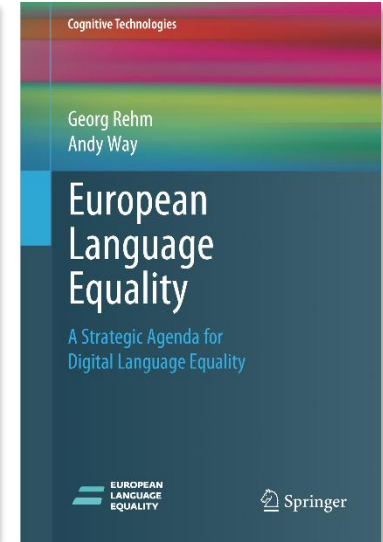
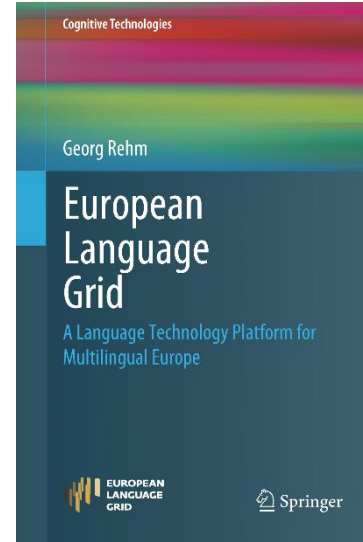
# Next Steps

- LDS is in full swing: technical development, promotion, dissemination, governance etc.
- Collaborations with
  - DSSC, Simpl and ALT-EDIC
  - European projects, e.g., HPLT, OpenGPT-X, OpenWebSearch
  - Other relevant data spaces, especially Media, Tourism and Cultural Heritage
  - EuroHPC
- Adoption of LDS by industry and other organisations [?](#) grow the LDS User Group
- Identify and make available new and fresh language data, especially from industry and covering all European languages and modalities



## Common European Language Data Space

Thank you!



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

Katrin Marheinecke Georg Rehm (DFKI GmbH, Germany) – LDS Project Coordination  
katrin.marheinecke@dfki.de

23-01-2025 LDS Country Workshop Belgium and Netherlands  
<https://language-data-space.ec.europa.eu>