EUROPEAN
LANGUAGE
DATA SPACE

Common European Language Data Space: Developing a market for language data and services and benefitting from a joint European effort

Firstname Lastname (Affiliation, Country)
name@domain.tld

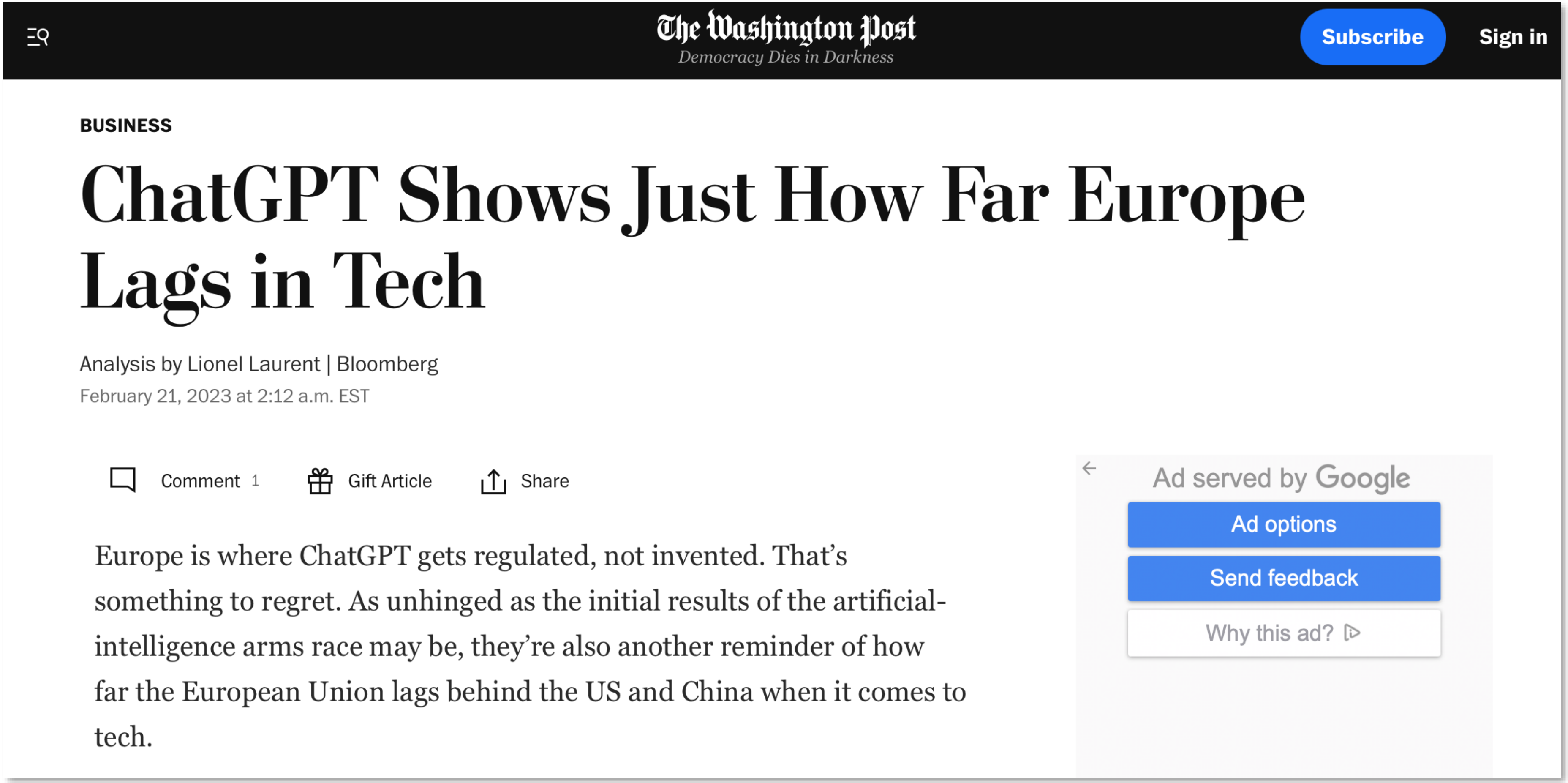
xx-xx-2024 LDS Country Workshop, City, Country
https://language-data-space.ec.europa.eu

EUROPEAN
LANGUAGE
DATA SPACE

Large Language Models

Context: Large Language Models

- Large language models are the most disruptive breakthrough in AI in recent history (BERT, GPT-3, ChatGPT, GPT-4 etc.)
- LLMs are based on vast amounts of training data
- LLMs use dozens, some even hundreds of terabytes of language and also image, video, audio etc. training data
- Europe’s languages are vastly under-resourced, except English
- A concerted effort for the collection of enormous amounts of language data for all European languages is very much needed
- Already now billions and billions are made but …



The Washington Post
Democracy Dies in Darkness

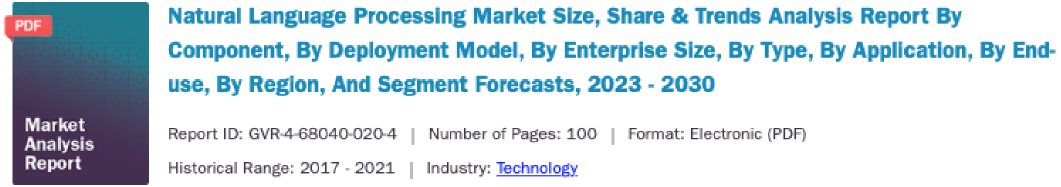
BUSINESS

ChatGPT Shows Just How Far Europe Lags in Tech

Analysis by Lionel Laurent | Bloomberg
February 21, 2023 at 2:12 a.m. EST

Europe is where ChatGPT gets regulated, not invented. That’s something to regret. As unhinged as the initial results of the artificial-intelligence arms race may be, they’re also another reminder of how far the European Union lags behind the US and China when it comes to tech.

Global LT/NLP Market is exploding: 439.85B$ by 2030



PDF

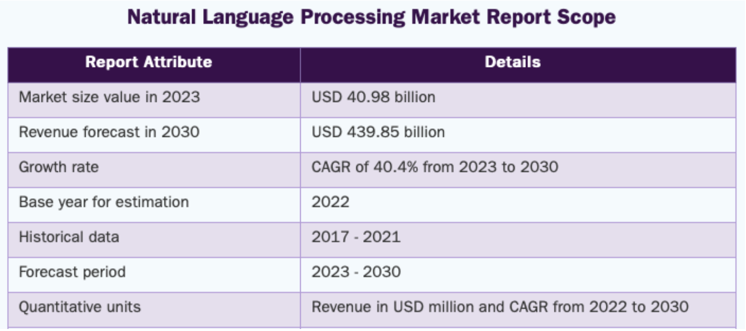
Market Analysis Report

Natural Language Processing Market Size, Share & Trends Analysis Report By Component, By Deployment Model, By Enterprise Size, By Type, By Application, By End-use, By Region, And Segment Forecasts, 2023 - 2030

Report ID: GVR-4-68040-020-4 | Number of Pages: 100 | Format: Electronic (PDF)

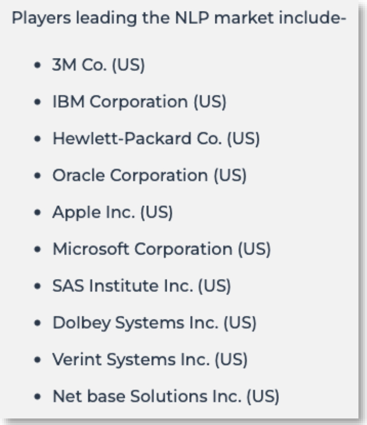
Historical Range: 2017 - 2021 | Industry: Technology

https://www.grandviewresearch.com/industry-analysis/natural-language-processing-market-report



Natural Language Processing Market Report Scope

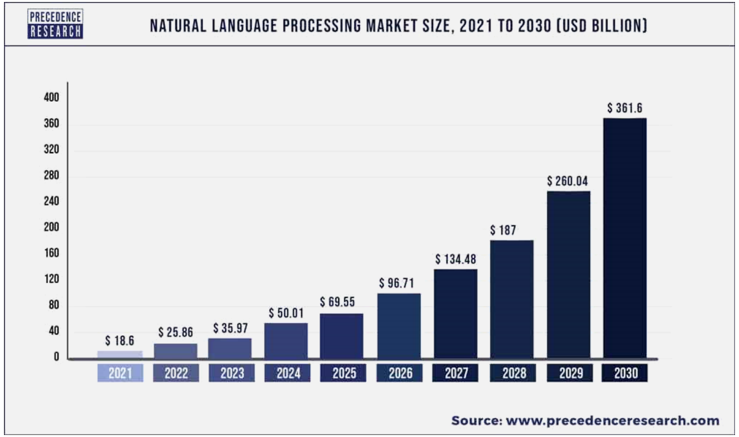
Report Attribute	Details
Market size value in 2023	USD 40.98 billion
Revenue forecast in 2030	USD 439.85 billion
Growth rate	CAGR of 40.4% from 2023 to 2030
Base year for estimation	2022
Historical data	2017 - 2021
Forecast period	2023 - 2030
Quantitative units	Revenue in USD million and CAGR from 2022 to 2030



Players leading the NLP market include-

- 3M Co. (US)
- IBM Corporation (US)
- Hewlett-Packard Co. (US)
- Oracle Corporation (US)
- Apple Inc. (US)
- Microsoft Corporation (US)
- SAS Institute Inc. (US)
- Dolbey Systems Inc. (US)
- Verint Systems Inc. (US)
- Net base Solutions Inc. (US)

All US!



Source: www.precedenceresearch.com

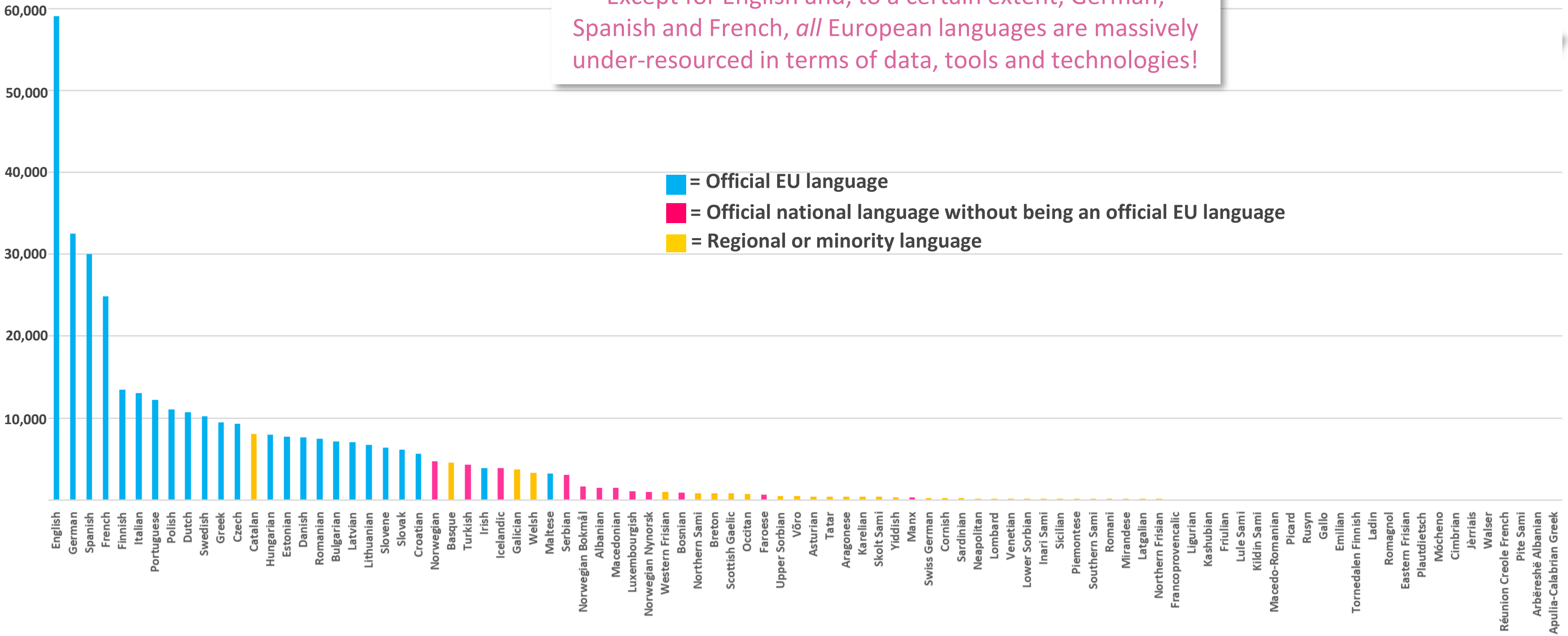
https://www.precedenceresearch.com/natural-language-processing-market

Without a decisive intervention by the EU, Europe will be pushed further to the side lines in the global NLP market.

Digital Language Equality?

EUROPEAN LANGUAGE EQUALITY

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!



EUROPEAN
LANGUAGE
DATA SPACE

Data Strategy & Data Spaces

EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy
- 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 approx. 20 official EU data space projects – focus on industry

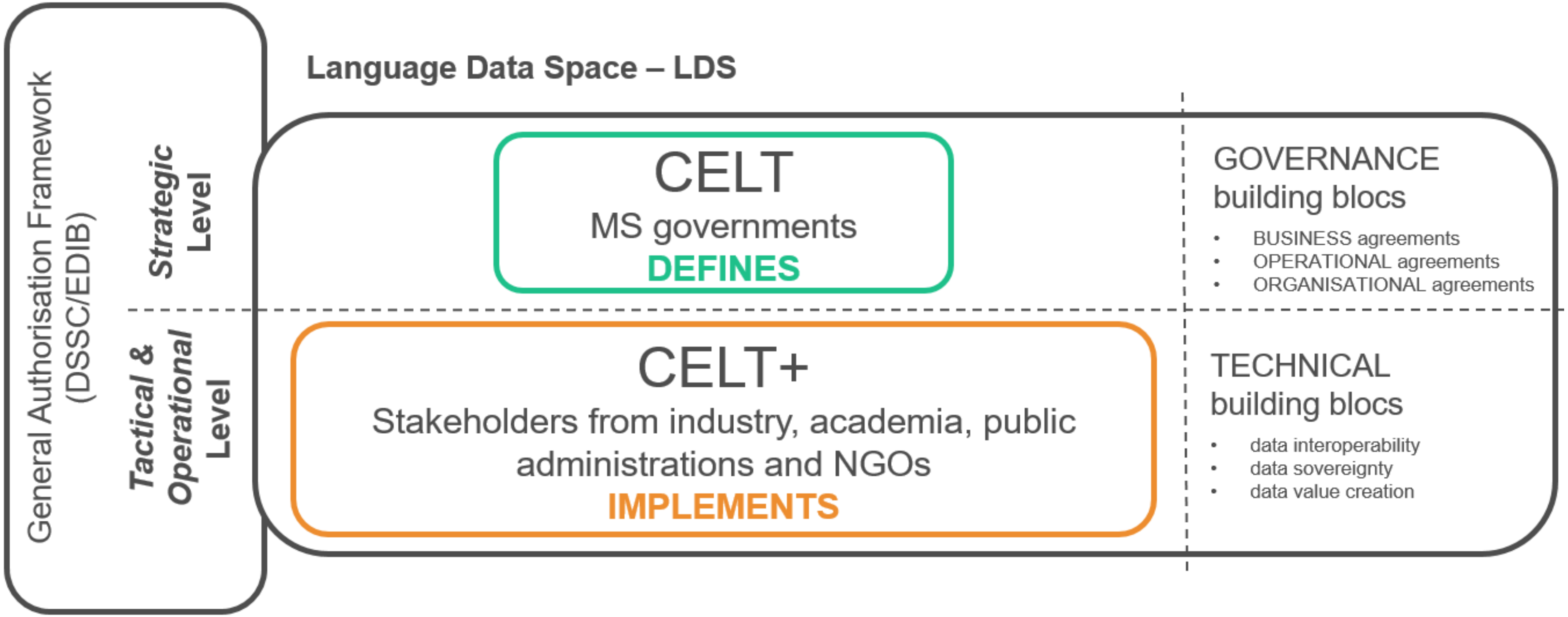
Language Data Space

- 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 architecture and infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens

Establishment of the CELT and CELT+

Establishment of the Centre of Excellence for Language Technologies (CELT)	Establishment of a Multi-Stakeholder Governance Body (CELT+)
Mission: strategic	Mission: tactical/operational
Members: governance representatives (from two different ministries from each Member State)	Members: companies, research centres and other organisations – also: identify other stakeholders (60%/40% private/public distribution of members)
Nucleus of the future Language EDIC	Periphery of the future Language EDIC
Address collection, creation, sharing and re-use of data and models; aggregate initiatives and coordinate LDS governance scheme	Coordinate development of the blueprint, focus upon the technical building blocks and operational aspects of the LDS
Operational four months after inception meeting	Operational six months after inception meeting
Statutes ready in Q4 2023	Statutes ready in Q4 2023/Q1 2024

CELT and CELT+ Governance Bodies



General Authorisation Framework (DSSC/EDIB)
Strategic Level
Tactical & Operational Level
Language Data Space – LDS
CELT
MS governments
DEFINES
CELT+
Stakeholders from industry, academia, public administrations and NGOs
IMPLEMENTS
GOVERNANCE building blocs
• BUSINESS agreements
• OPERATIONAL agreements
• ORGANISATIONAL agreements
TECHNICAL building blocs
• data interoperability
• data sovereignty
• data value creation

Previous Projects and Initiatives

- The four core partners – DFKI, ILSP, ELDA, TILDE – have been involved in many joint projects and initiatives, 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)

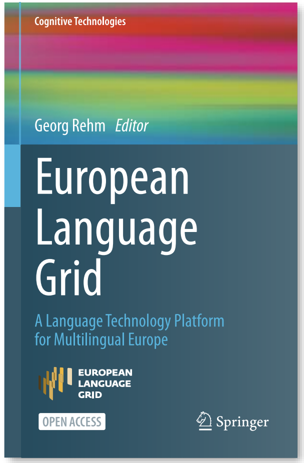
META-NET

European Language Resource Coordination
Connecting Europe Facility

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY

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



Cognitive Technologies

Georg Rehm *Editor*

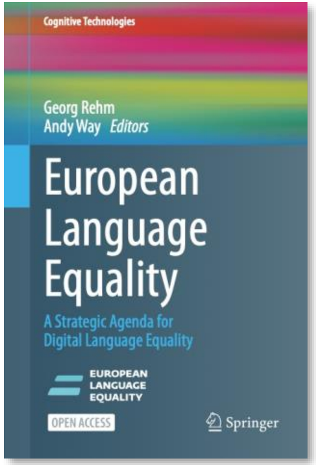
European Language Grid

A Language Technology Platform for Multilingual Europe

EUROPEAN LANGUAGE GRID

OPEN ACCESS

Springer



Cognitive Technologies

Georg Rehm
Andy Way *Editors*

European Language Equality

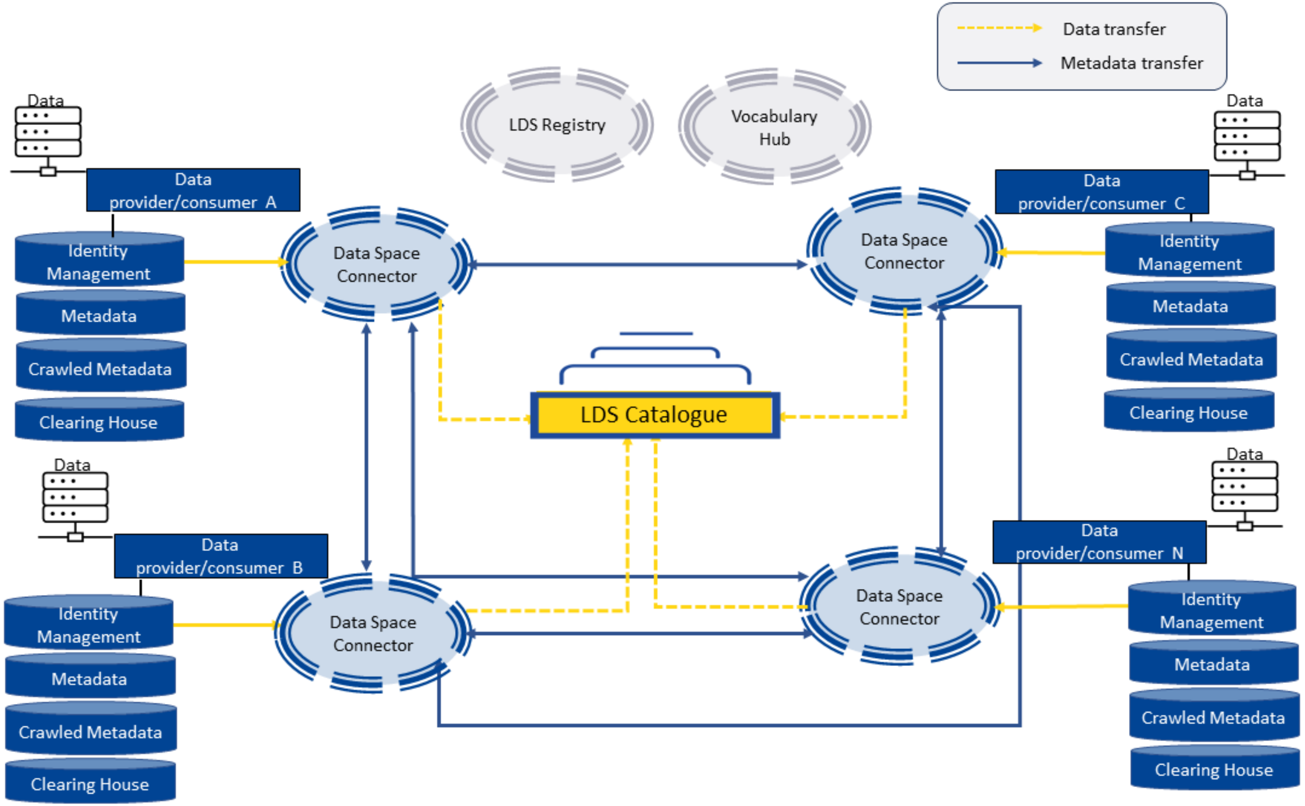
A Strategic Agenda for Digital Language Equality

EUROPEAN LANGUAGE EQUALITY

OPEN ACCESS

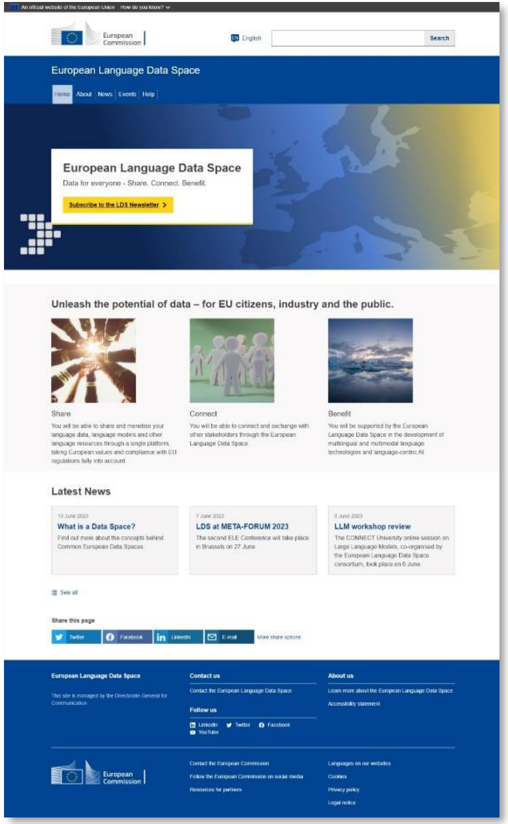
Springer

The LDS Decentralised Architecture



Collaborations (selection)

- **DSSC** (Digital, EU; Community of Practice; Thematic Groups; Expert Groups)
- **OpenGPT-X** (Gaia-X; BMWK, Germany)
- **HPLT** (EU Horizon Europe)
- **DataBri-X** (IDSA; EU Horizon Europe)
- **European Language Grid** (ELG) – currently supported through OpenGPT-X, SciLake, DataBri-X – legal entity work in progress
- **European Language Equality** (ELE, EU PP/PA project)
- **INESData** (new language data space project in Spain; 65% of the 5M€ funding for industry for development of actual platform)
- **SciLake** (EOSC; EU Horizon Europe)



An official website of the European Union How do you know?

European Commission

English Search

European Language Data Space

Home About News Events Help

European Language Data Space

Data for everyone - Share. Connect. Benefit.

Subscribe to the LDS Newsletter >

Unleash the potential of data – for EU citizens, industry and the public.

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

Latest News

13 June 2023
What is a Data Space?
Find out more about the concepts behind Common European Data Spaces

7 June 2023
LDS at META-FORUM 2023
The second ELE Conference will take place in Brussels on 27 June

5 June 2023
LLM workshop review
The CONNECT University online session on Large Language Models, co-organised by the European Language Data Space consortium, took place on 6 June

See all

Share this page

Twitter Facebook LinkedIn E-mail More share options

European Language Data Space

This site is managed by the Directorate-General for Communication

Contact us

Contact the European Language Data Space

Follow us

LinkedIn Twitter Facebook YouTube

About us

Learn more about the European Language Data Space

Accessibility statement

European Commission

Contact the European Commission

Follow the European Commission on social media

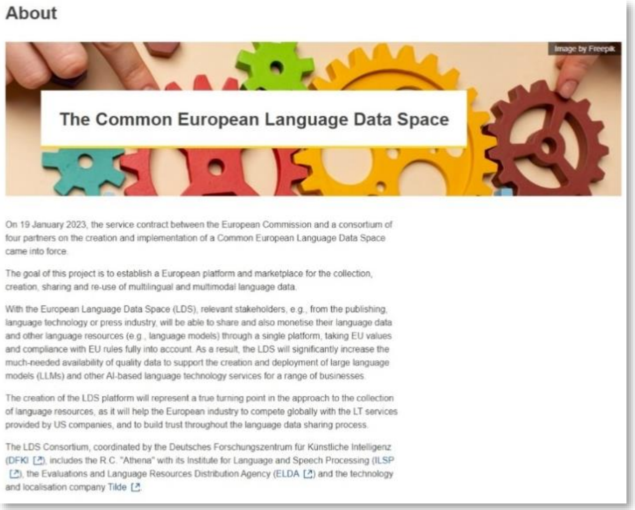
Resources for partners

Languages on our websites

Cookies

Privacy policy

Legal notice



About

Image by Freepik

The Common European Language Data Space

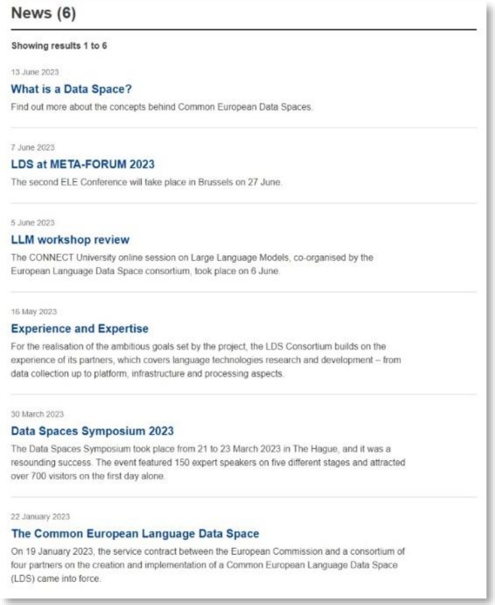
On 19 January 2023, the service contract between the European Commission and a consortium of four partners on the creation and implementation of a Common European Language Data Space came into force.

The goal of this project is to establish a European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data.

With the European Language Data Space (LDS), relevant stakeholders, e.g., from the publishing, language technology or press industry, will be able to share and also monetise their language data and other language resources (e.g., language models) through a single platform, taking EU values and compliance with EU rules fully into account. As a result, the LDS will significantly increase the much-needed availability of quality data to support the creation and deployment of large language models (LLMs) and other AI-based language technology services for a range of businesses.

The creation of the LDS platform will represent a true turning point in the approach to the collection of language resources, as it will help the European industry to compete globally with the LT services provided by US companies, and to build trust throughout the language data sharing process.

The LDS Consortium, coordinated by the Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI ↗), includes the R.C. "Athena" with its Institute for Language and Speech Processing (ILSP ↗), the Evaluations and Language Resources Distribution Agency (ELDA ↗) and the technology and localisation company Tilde ↗.



News (6)

Showing results 1 to 6

13 June 2023
What is a Data Space?
Find out more about the concepts behind Common European Data Spaces.

7 June 2023
LDS at META-FORUM 2023
The second ELE Conference will take place in Brussels on 27 June.

5 June 2023
LLM workshop review
The CONNECT University online session on Large Language Models, co-organised by the European Language Data Space consortium, took place on 6 June.

16 May 2023
Experience and Expertise
For the realisation of the ambitious goals set by the project, the LDS Consortium builds on the experience of its partners, which covers language technologies research and development – from data collection up to platform, infrastructure and processing aspects.

30 March 2023
Data Spaces Symposium 2023
The Data Spaces Symposium took place from 21 to 23 March 2023 in The Hague, and it was a resounding success. The event featured 150 expert speakers on five different stages and attracted over 700 visitors on the first day alone.

22 January 2023
The Common European Language Data Space
On 19 January 2023, the service contract between the European Commission and a consortium of four partners on the creation and implementation of a Common European Language Data Space (LDS) came into force.

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

Common European Language Data Space

Thank you!

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

Firstname Lastname (Affiliation, Country)
name@domain.tld

xx-xx-2023 Name of the event, City, Country
https://language-data-space.ec.europa.eu