EUROPEAN
LANGUAGE
DATA SPACE

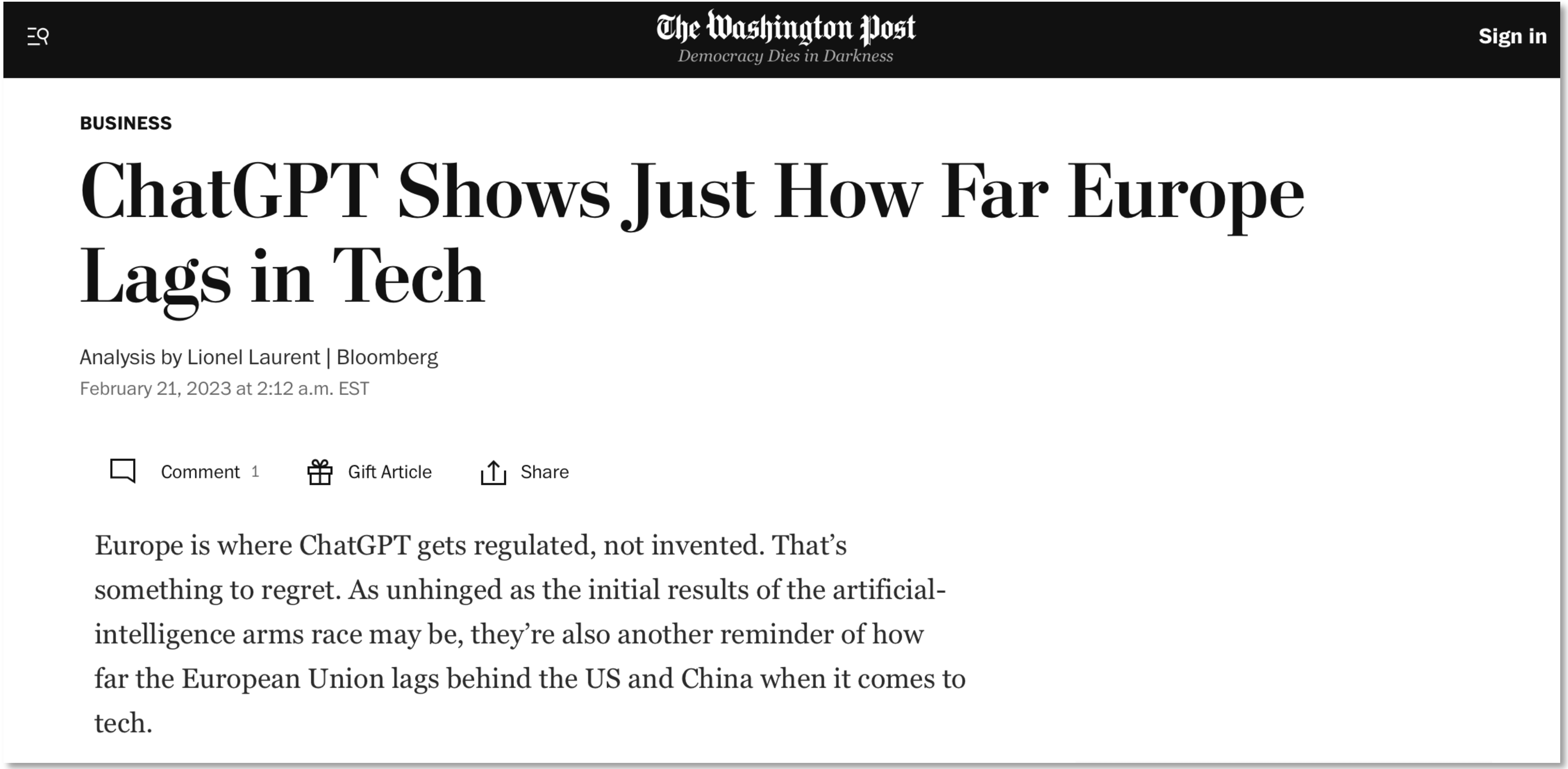
European Language Data Space

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
georg.rehm@dfki.de

29-01-2024 2nd Technology Workshop
https://language-data-space.ec.europa.eu

Context: Large Language Models (LLMs)

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



The Washington Post

Democracy Dies in Darkness

BUSINESS

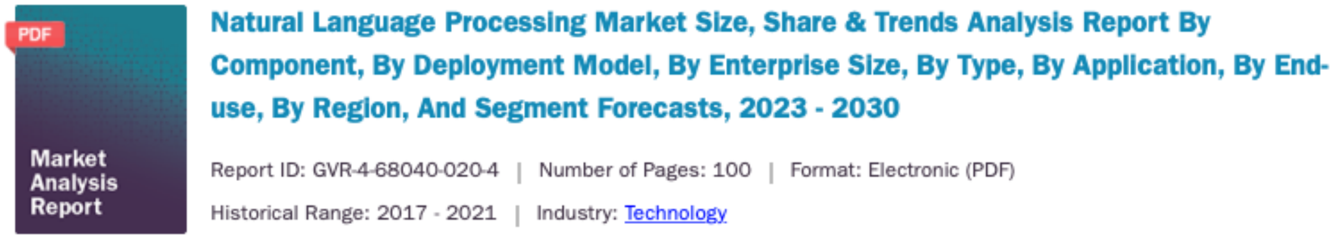
ChatGPT Shows Just How Far Europe Lags in Tech

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

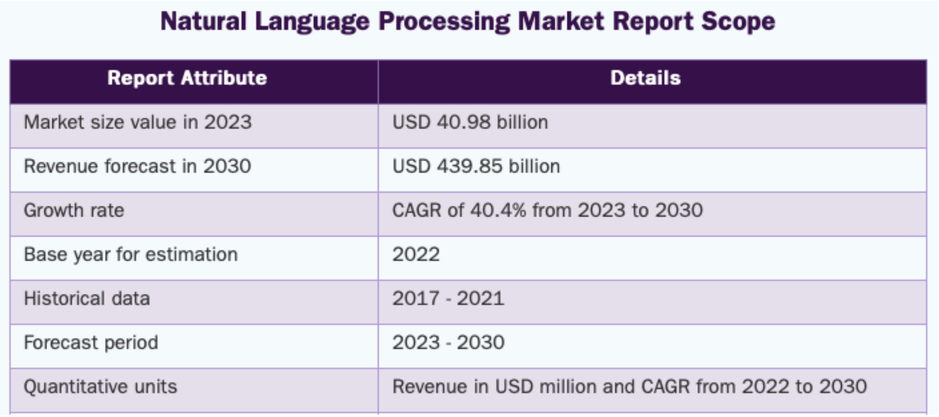
Market Analysis Report

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)

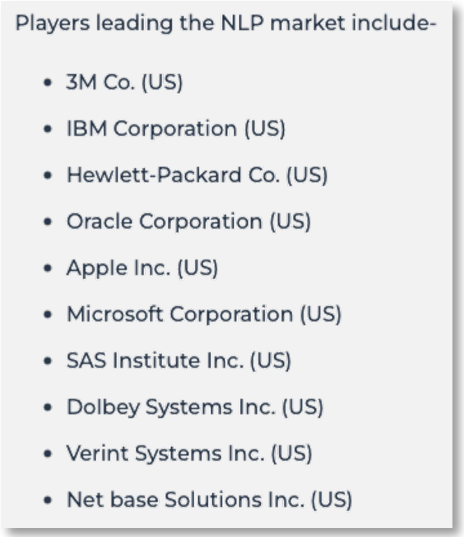
Historical Range: 2017 - 2021 | Industry: Technology

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



Natural Language Processing Market Report Scope

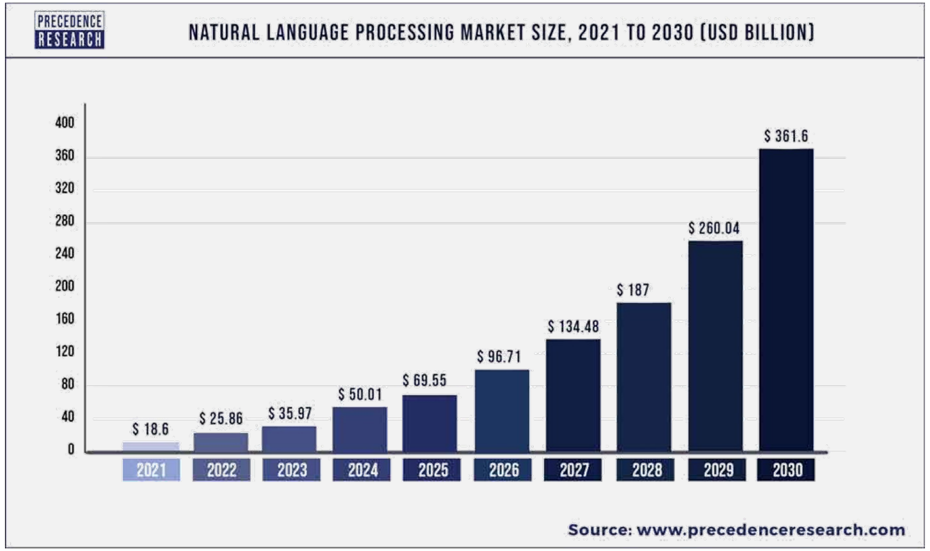
Report Attribute	Details
Market size value in 2023	USD 40.98 billion
Revenue forecast in 2030	USD 439.85 billion
Growth rate	CAGR of 40.4% from 2023 to 2030
Base year for estimation	2022
Historical data	2017 - 2021
Forecast period	2023 - 2030
Quantitative units	Revenue in USD million and CAGR from 2022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

EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy and key instrument for the data economy
- The EU project DSSC is currently helping to develop the technical and operational specifications
  - IDSA, Gaia-X, FI-WARE, BDVA are members of DSSC – convergence of data space approaches!
- Important goals and properties of data spaces:
  - Technical sovereignty and data sovereignty
  - Establishing trust between participants
  - Data transactions can be free or for a certain price
  - Data providers/owners will be fully in control of their own data
  - Policies, standards, rules, contracts will be technically enforced
- The LDS is one of the currently 14 official EU data space projects – 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, selling 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

EUROPEAN
LANGUAGE
DATA SPACE

Consortium and Subcontractors

Lead Partner and Coordinator		
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	DE
Partners and Operation Leads		
R.C. “Athena”, Institute for Language and Speech Processing	ILSP	GR
Evaluations and Language Resources Distribution Agency	ELDA	FR
TILDE	TILDE	LV
Main Subcontractors		
3pc GmbH Neue Kommunikation	3pc	DE
Capgemini Deutschland GmbH	CapG	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

Centre of Excellence for Language Technologies (CELT)	Multi-Stakeholder Governance Body: LDS User Group
Mission: strategic	Mission: tactical/operational
Members: government representatives (from up to 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)
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

Typical Use Cases of the LDS as a Data Space and Data Marketplace

Organisations that develop NLP/LLM technologies

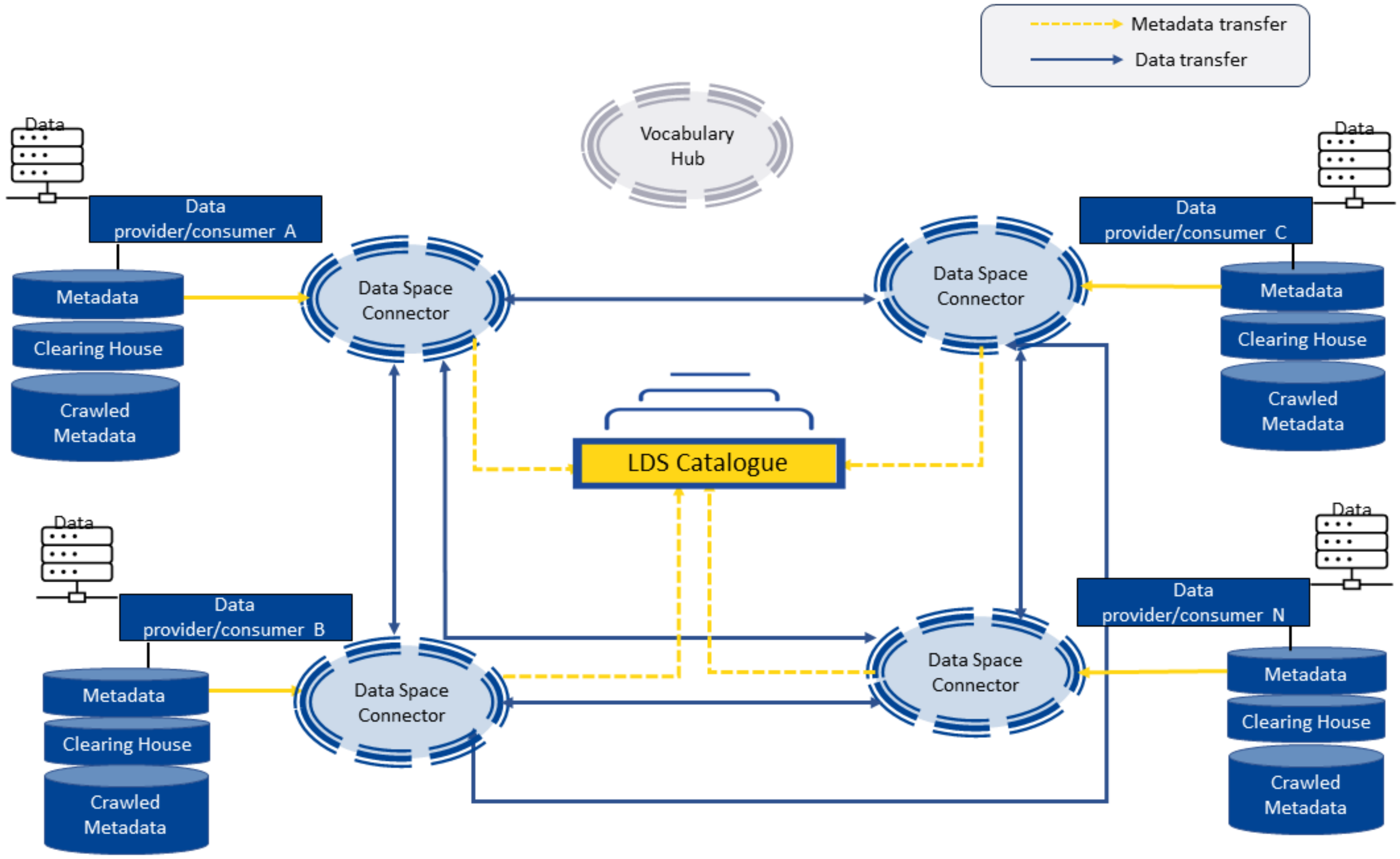
- Check if data sets are available in LDS that cover a certain language and purchase as well as download those that address their need
- Check if data sets are available in LDS that cover a certain language and genre/register and purchase as well as download those that address their need

Organisations that would like to provide language data

- Make language data sets available, either for free or for a certain price; these organisations will *always* remain in control of their own data; terms of use can be based on templates or bespoke

Data preprocessing through organisations that offer certain services

- Certain data sets may need preprocessing before they are made available through LDS; such preprocessing services, e.g., anonymisation, can also be offered through LDS


Metadata transfer
Data transfer
Vocabulary Hub
Data
Data provider/consumer A
Metadata
Clearing House
Crawled Metadata
Data Space Connector
Data
Data provider/consumer C
Metadata
Clearing House
Crawled Metadata
Data Space Connector
LDS Catalogue
Data
Data provider/consumer B
Metadata
Clearing House
Crawled Metadata
Data Space Connector
Data
Data provider/consumer N
Metadata
Clearing House
Crawled Metadata
Data Space Connector

Large Language Models *for* Europe *made in* Europe – The Big Picture

LLM Platform

- User-friendly platform for pre-training, fine-tuning and deploying LLMs
- Can be realised on top of ELG (requires additional development work & HPC access)
- Huge gap that needs to be addressed soon

EUROPEAN LANGUAGE GRID ?

Data Curation

- High-performance tools for language data curation, filtering and annotation
- Filtering, quality and bias assessment, identification and analysis of various text and document properties
- Huge gap that needs to be addressed soon

?

Federated Learning

- Federated learning of LLMs by making use of the data sets provided in a decentralised way by LDS
- Huge gap that needs to be addressed soon

?

NLP/LT Platform

- General NLP and Language Technology Platform
- Provide general, i.e., non-LLM-based, NLP and LT services (e.g., anonymisation)
- Addressed by European Language Grid

EUROPEAN LANGUAGE GRID

Compute

- High-Performance Compute infrastructure for training LLMs
- Addressed by EuroHPC Joint Undertaking
- Simple yet crucial issue: storage capacity, esp. for multimodal data

EuroHPC Joint Undertaking

Language Data

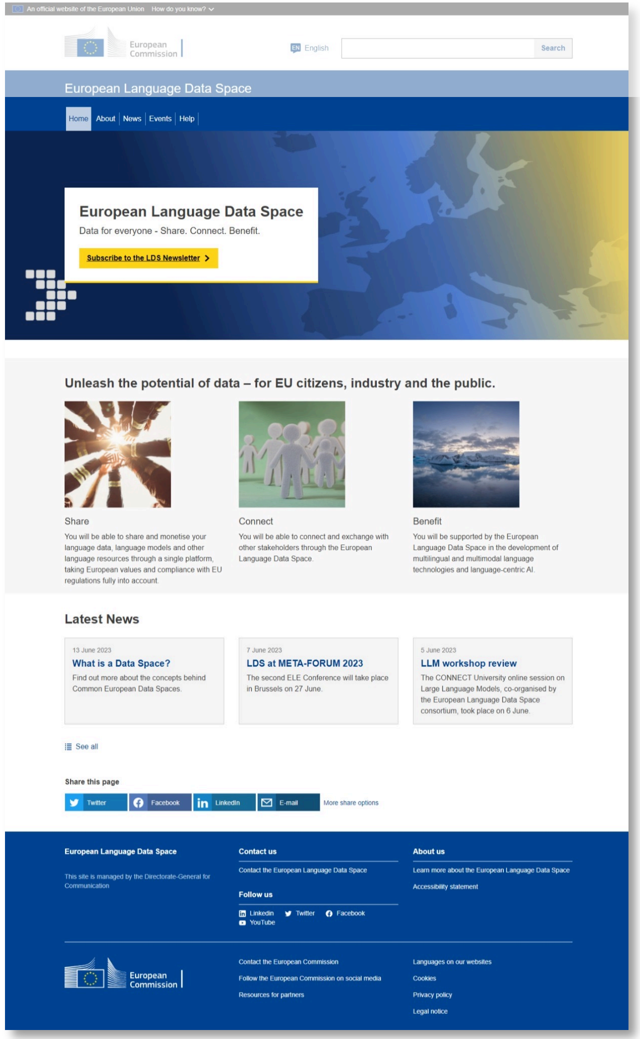
- Secure and trusted sharing of language data covering all European languages
- Decentralised, i.e., federated
- Addressed by Language Data Space

EUROPEAN LANGUAGE DATA SPACE

Next Steps

- LDS is in full swing: technical development, promotion, dissemination, governance etc.
- Collaboration with DSSC and collaboration with ALT-EDIC Working Group
- Collaboration with LLM projects such as HPLT and OpenGPT-X
- Collaboration with EuroHPC JU
- Collaboration with other data spaces, especially Media and Cultural Heritage
- Very important next step:
  - Raising awareness and adoption of LDS by industry and other organisations
  - Identify and make available new and fresh language data, especially from industry and ideally covering all European languages and modalities

Subscribe to our newsletter!



European Language Data Space

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

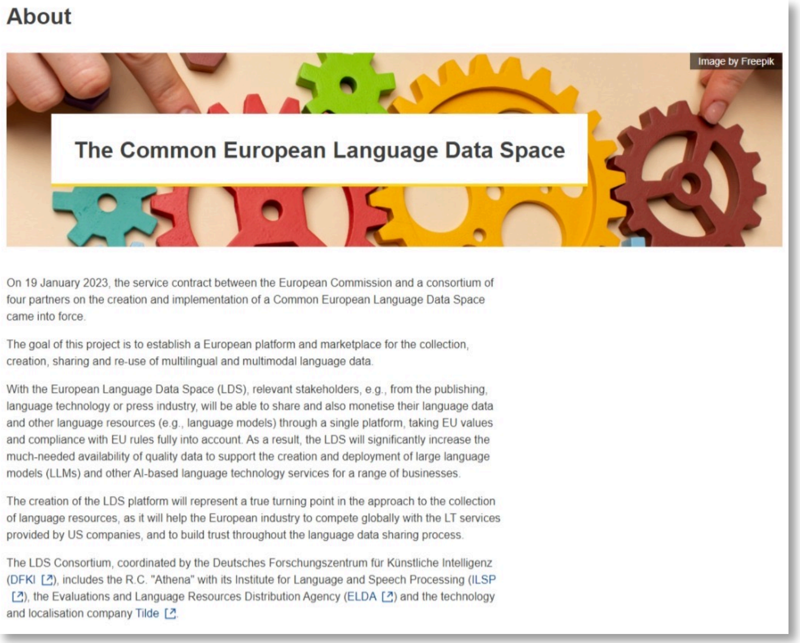
Resources for partners

Languages on our websites

Cookies

Privacy policy

Legal notice



About

The Common European Language Data Space

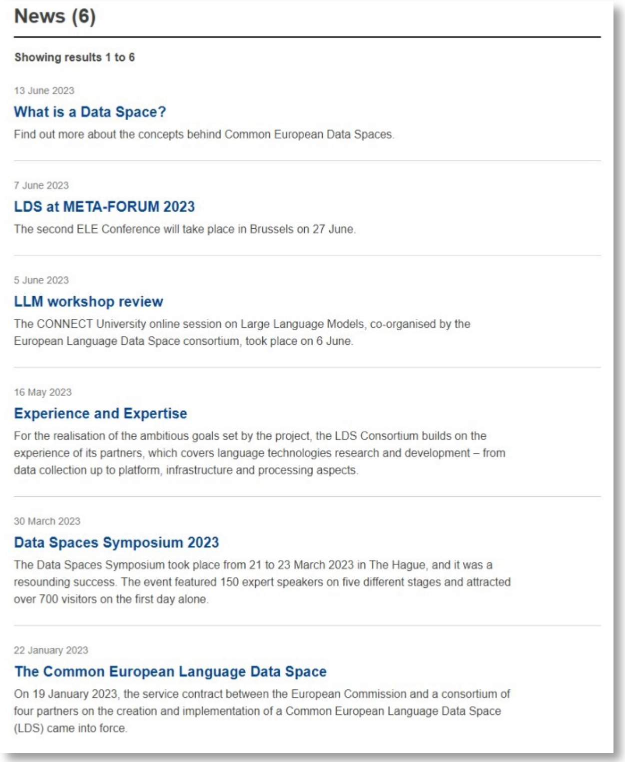
On 19 January 2023, the service contract between the European Commission and a consortium of four partners on the creation and implementation of a Common European Language Data Space came into force.

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.

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
georg.rehm@dfki.de

29-01-2024 2nd Technology Workshop
https://language-data-space.ec.europa.eu