EUROPEAN
LANGUAGE
EQUALITY

META-NET
META-FORUM 2023

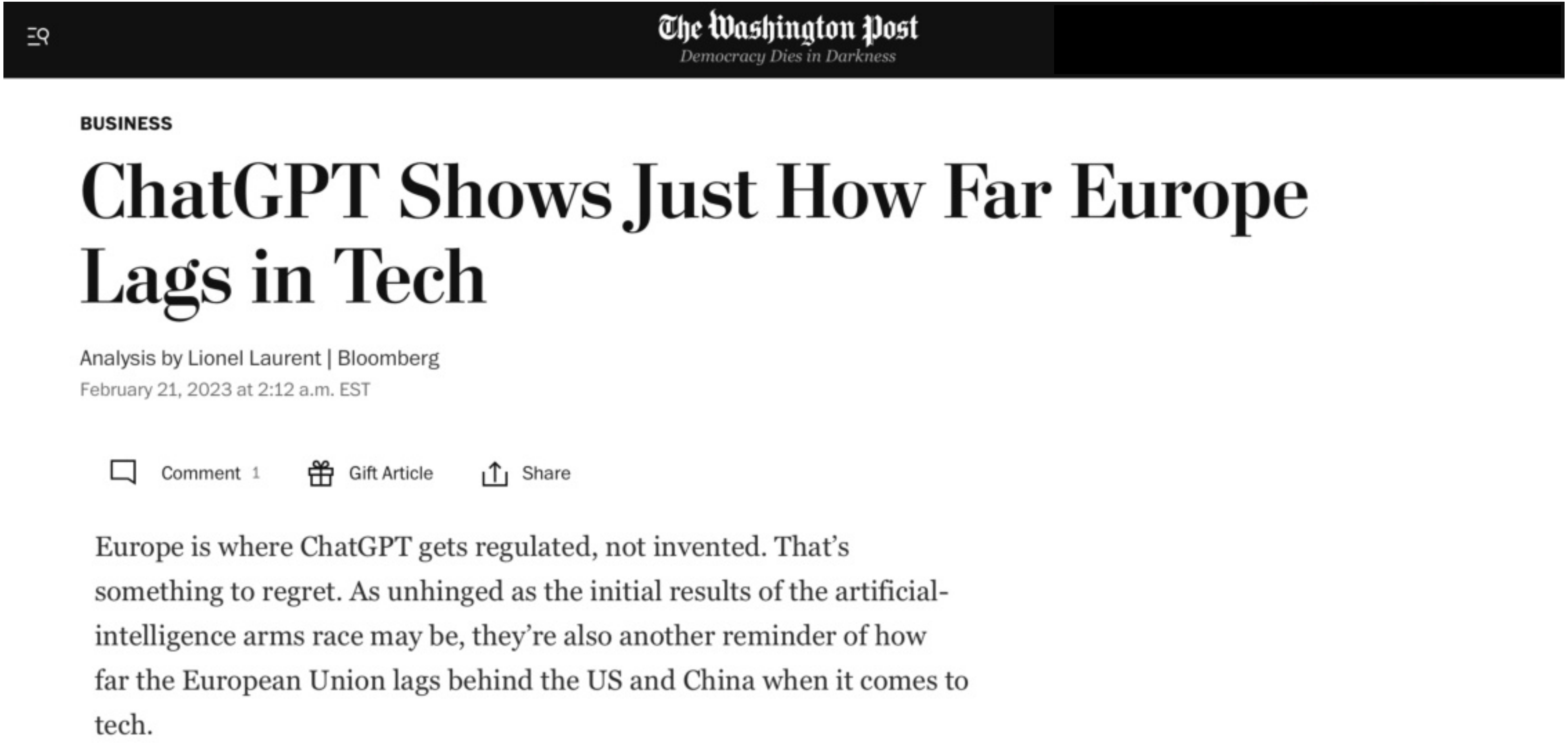
Language Data Space

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

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

Context: Large Language Models (LLMs)

- Large language models are the most disruptive breakthrough in AI in the last 15-20 years (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 (billions of tokens) and image, video, audio etc. data
- Europe's languages are *vastly* under-resourced, except English
- A concerted effort for the collection of vast 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

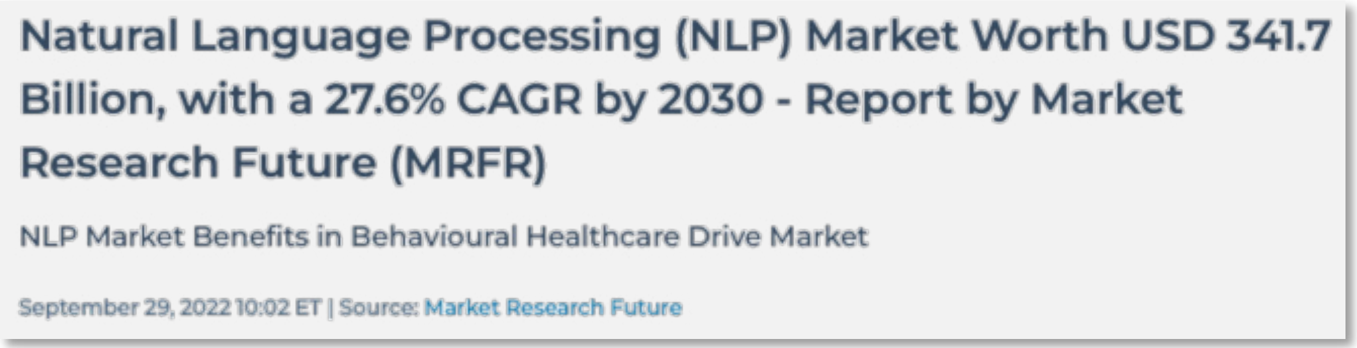
BUSINESS

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 and NLP Market by 2030



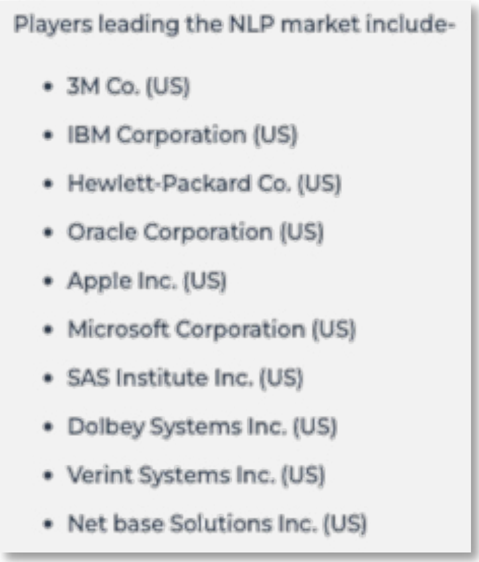
Natural Language Processing (NLP) Market Worth USD 341.7 Billion, with a 27.6% CAGR by 2030 - Report by Market Research Future (MRFR)

NLP Market Benefits in Behavioural Healthcare Drive Market

September 29, 2022 10:02 ET | Source: Market Research Future

Note: *these numbers are pre-ChatGPT!*

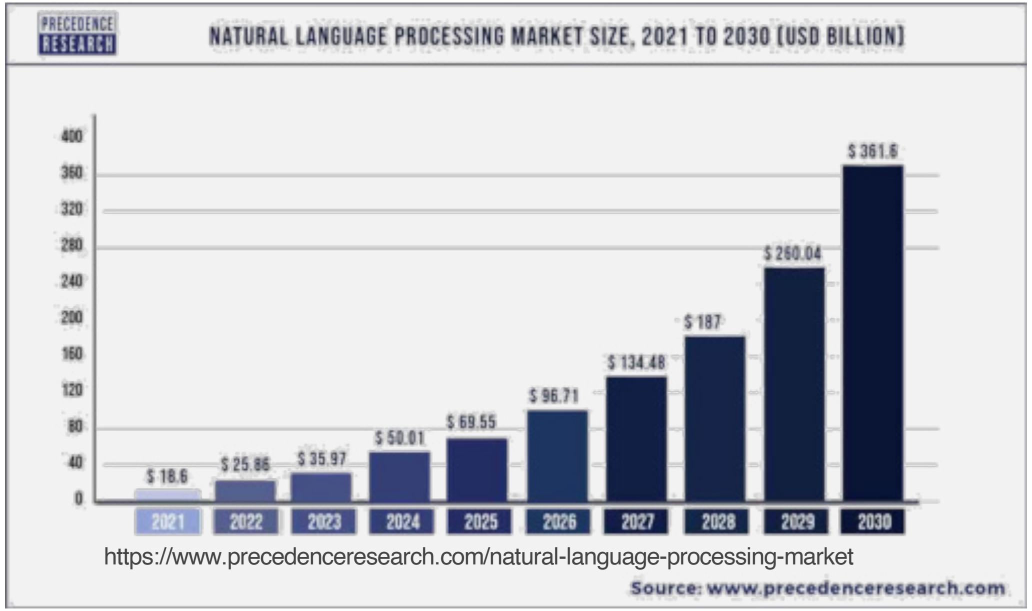
https://www.globenewswire.com/en/news-release/2022/09/29/2525379/0/en/Natural-Language-Processing-NLP-Market-Worth-USD-341-7-Billion-with-a-27-6-CAGR-by-2030-Report-by-Market-Research-Future-MRFR.html



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!



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

Source: www.precedenceresearch.com

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

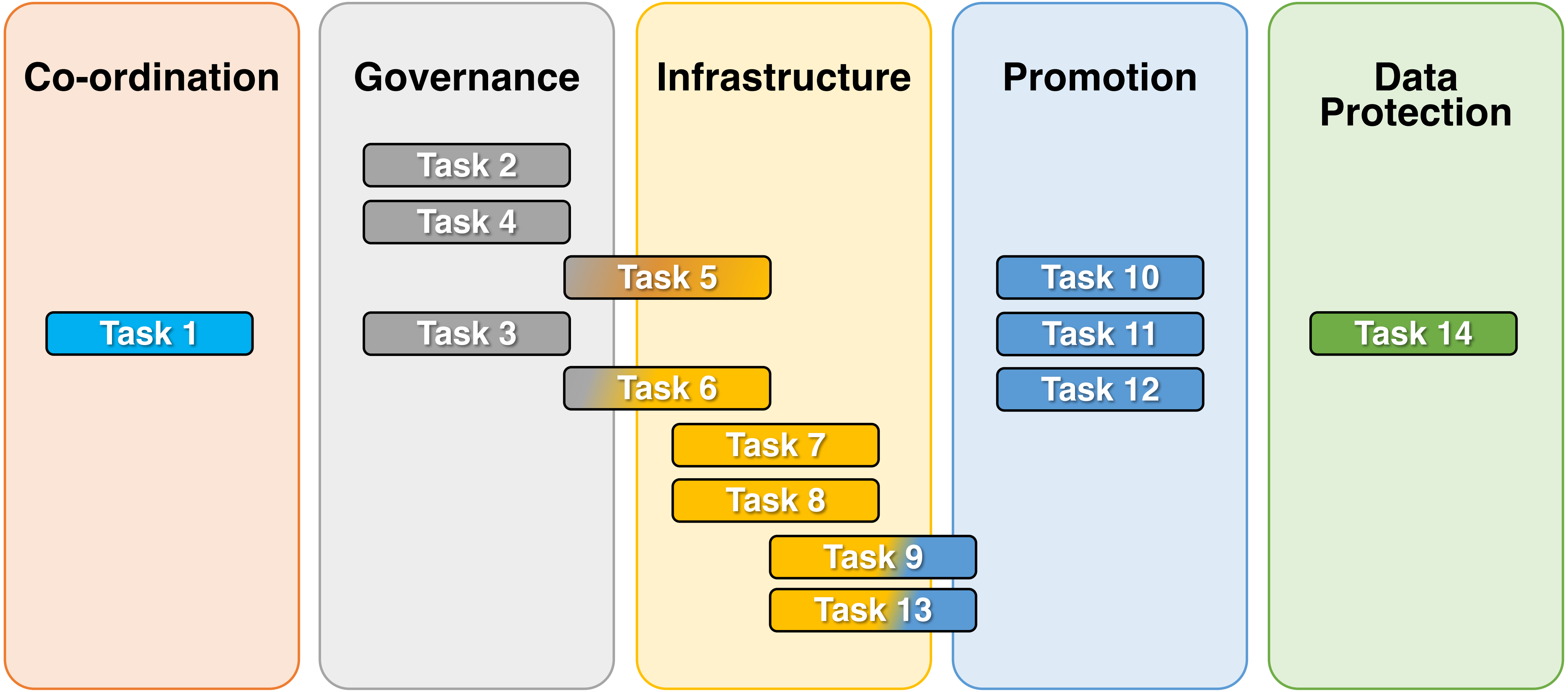
- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget/length: 6M€ + 2M€. 36 + 12 months (if renewed)
- Objective: Develop and deploy a European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- Salient features: governance framework, technical architecture and infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens

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
Data, AI and Robotics AISBL	BDVA	BE

plus legal experts (Delcade) and approx. 30 organisations to support LDS with the organisation of country workshops

Tasks



Co-ordination

Task 1

Governance

Task 2

Task 4

Task 3

Infrastructure

Task 5

Task 6

Task 7

Task 8

Task 9

Task 13

Promotion

Task 10

Task 11

Task 12

Data Protection

Task 14

Tasks

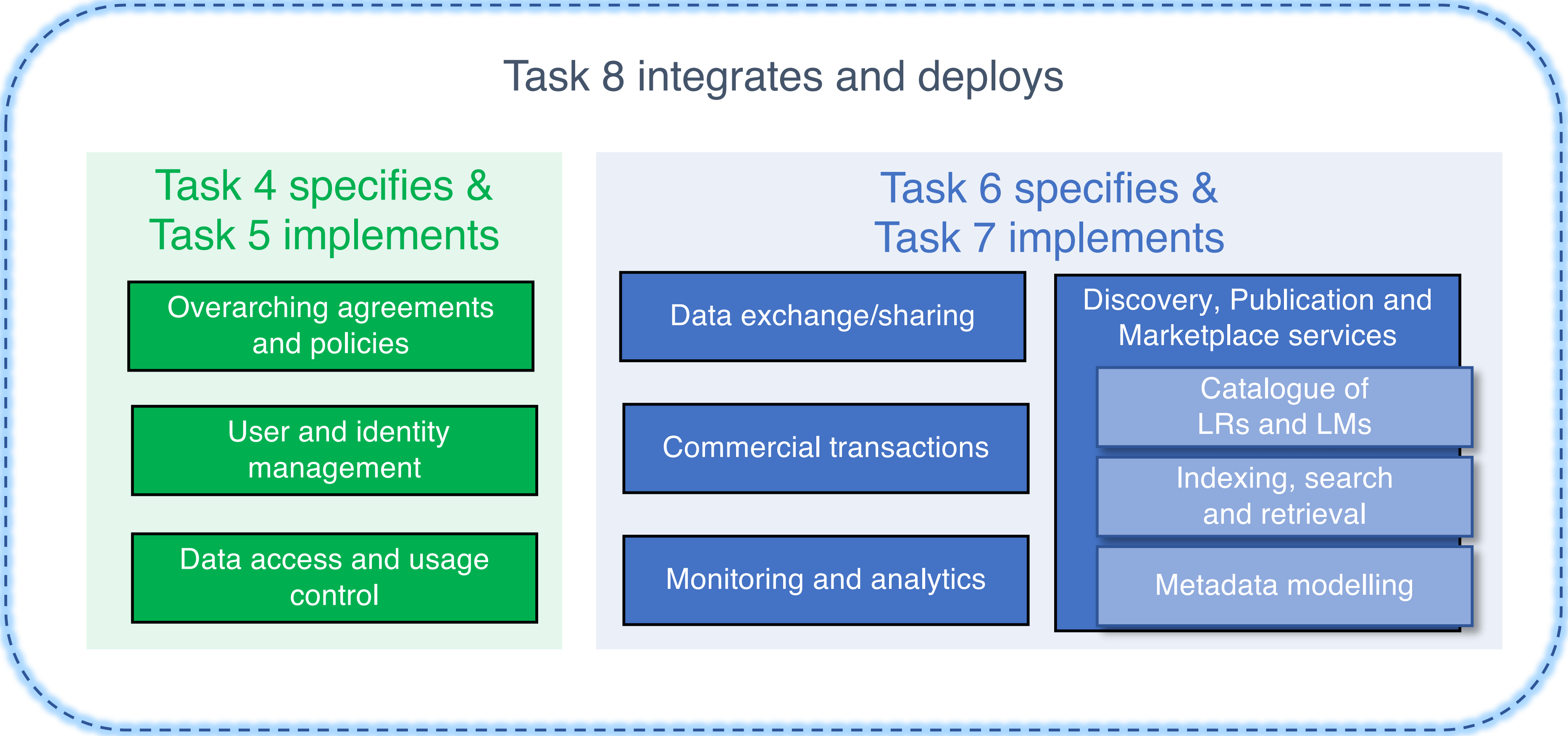
	Task	Lead
T1	Coordination and support	DFKI
T2	Establishment of the CELT	DFKI
T3	Establishment of the CELT+	DFKI
T4	Development of a Multi-Stakeholder Data and Services Governance Scheme	ELDA
T5	Implementation of Multi-Stakeholder Data and Services Governance Scheme	TILDE
T6	Development of a Sustainable Language Data Ecosystem Blueprint	ILSP
T7	Implementation of the Sustainable Language Data Ecosystem Blueprint	ILSP
T8	Language Data Space Deployment	ILSP
T9	Proof-of-Deployment-Concept Projects	TILDE
T10	Event Organisation and Management	ELDA
T11	Promotional Activities through Conference Attendance	DFKI
T12	Promotional Activities through Information Channels	DFKI
T13	Language Data Space Website	DFKI
T14	Data Protection Compliance	ELDA

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, EU Network of Excellence, 2010-2013)
  - META-SHARE
- **ELRC** (various CEF projects/contracts, 2014-2023)
  - ELRC-SHARE
- **ELG** (Horizon 2020, 2019-2022)
  - ELG Cloud Platform
- **ELE** (PP/PA, 2021-2022; 2022-2023)

The **technical development** of the LDS platform will be informed by experiences made with ELG, ELRC-SHARE, META-SHARE and others.

LDS Technical Components



Task 8 integrates and deploys

Task 4 specifies &
Task 5 implements

Overarching agreements and policies

User and identity management

Data access and usage control

Task 6 specifies &
Task 7 implements

Data exchange/sharing

Commercial transactions

Monitoring and analytics

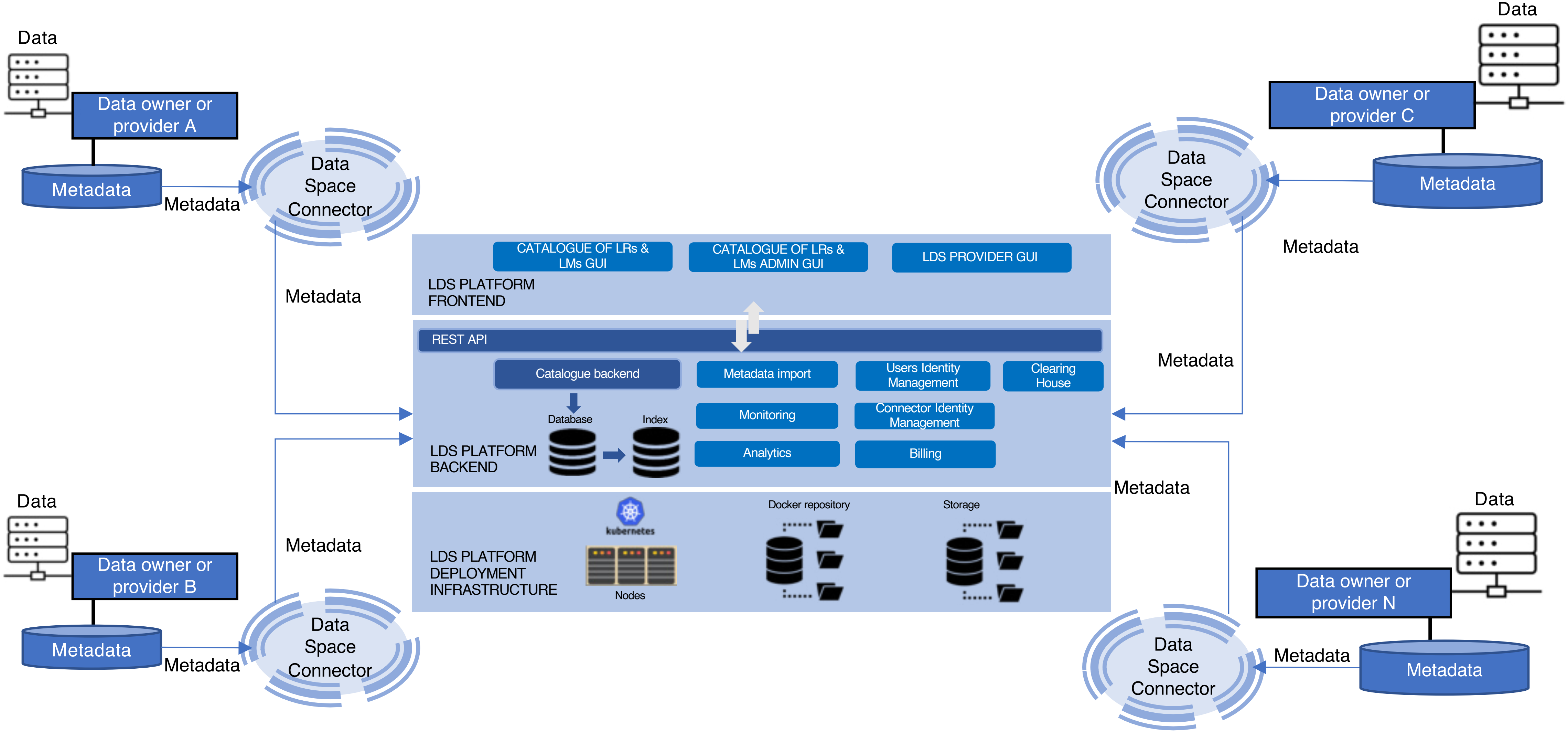
Discovery, Publication and Marketplace services

Catalogue of LRs and LMs

Indexing, search and retrieval

Metadata modelling

Possible Architecture



Data

Data owner or provider A

Metadata

Metadata

Data Space Connector

Metadata

Data

Data owner or provider B

Metadata

Metadata

Data Space Connector

Metadata

Data

Data owner or provider C

Metadata

Data Space Connector

Metadata

Metadata

Data

Data owner or provider N

Metadata

Metadata

Data Space Connector

Metadata

CATALOGUE OF LRs & LMs GUI

CATALOGUE OF LRs & LMs ADMIN GUI

LDS PROVIDER GUI

LDS PLATFORM FRONTEND

REST API

Catalogue backend

Metadata import

Users Identity Management

Clearing House

Database

Index

Monitoring

Connector Identity Management

LDS PLATFORM BACKEND

Analytics

Billing

kubernetes

Docker repository

Storage

LDS PLATFORM DEPLOYMENT INFRASTRUCTURE

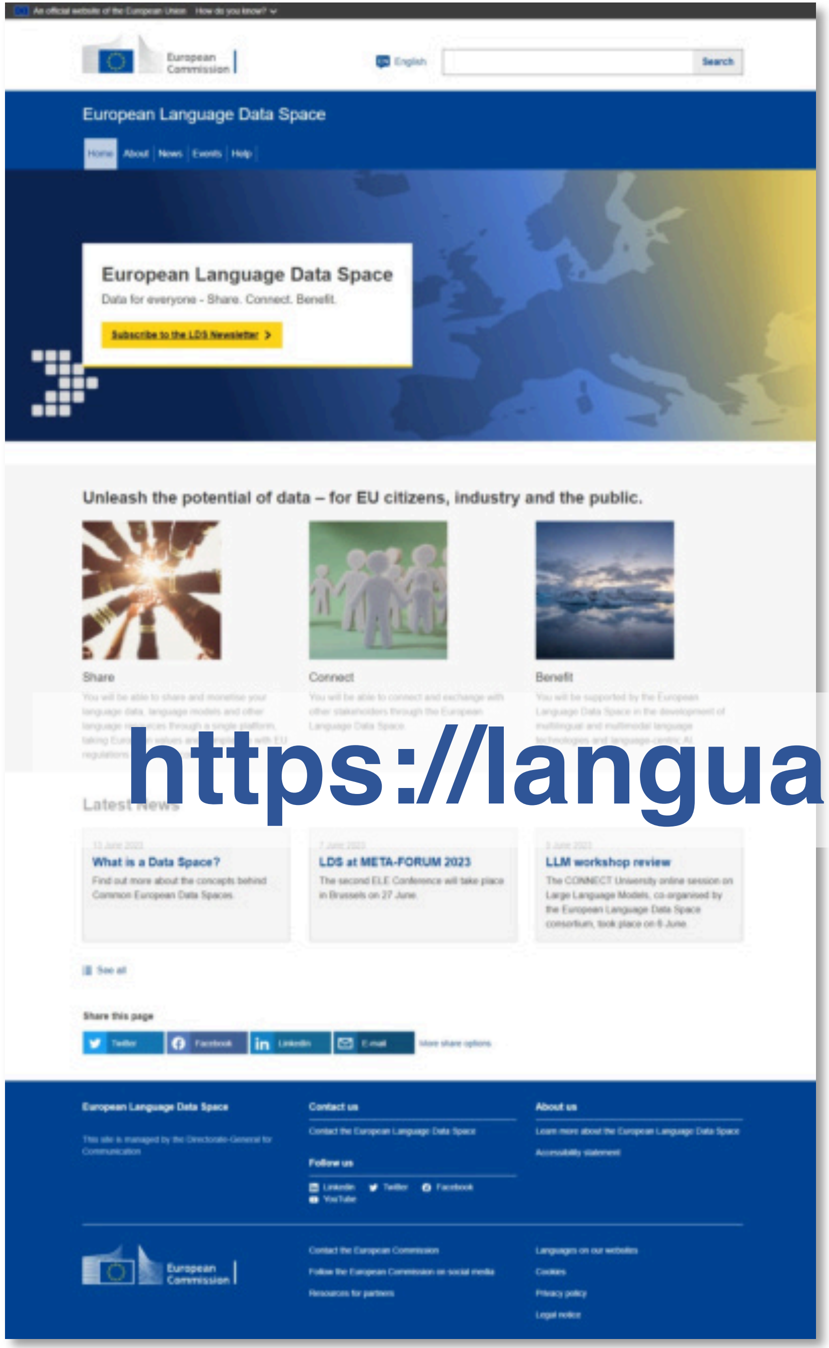
Nodes

Collaborations (selection)

- **DSSC** (Digital, EU; Community of Practice; The Hague event)
- **OpenGPT-X** (Gaia-X; BMWK, Germany)
- **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 the platform)
- **SciLake** (EOSC; EU Horizon Europe)

Next Steps (selection)

- Prepare LDS website on europa.eu – *work in progress*
- Make basic technical decisions (based on previous work, i.e., META-SHARE, ELRC, ELG, ELE) – *work in progress*
- Populate the main boards (CELT, CELT+) – *work in progress*
- Event planning (approx. 70!) – *work in progress*
- Collaboration with Language EDIC WG – *work in progress*



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

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

European Language Data Space

This site is managed by the Directorate-General for Communication

Contact us

Contact the European Language Data Space

Follow us

About us

Learn more about the European Language Data Space

Accessibility statement

Contact the European Commission

Follow the European Commission on social media

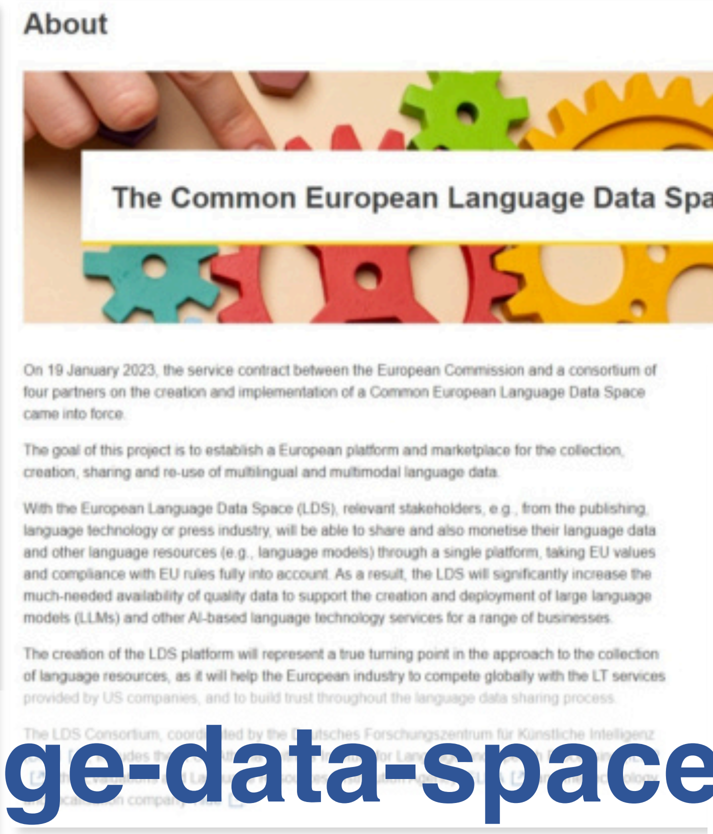
Resources for partners

Languages on our websites

Cookies

Privacy policy

Legal notice



About

The Common European Language Data Space

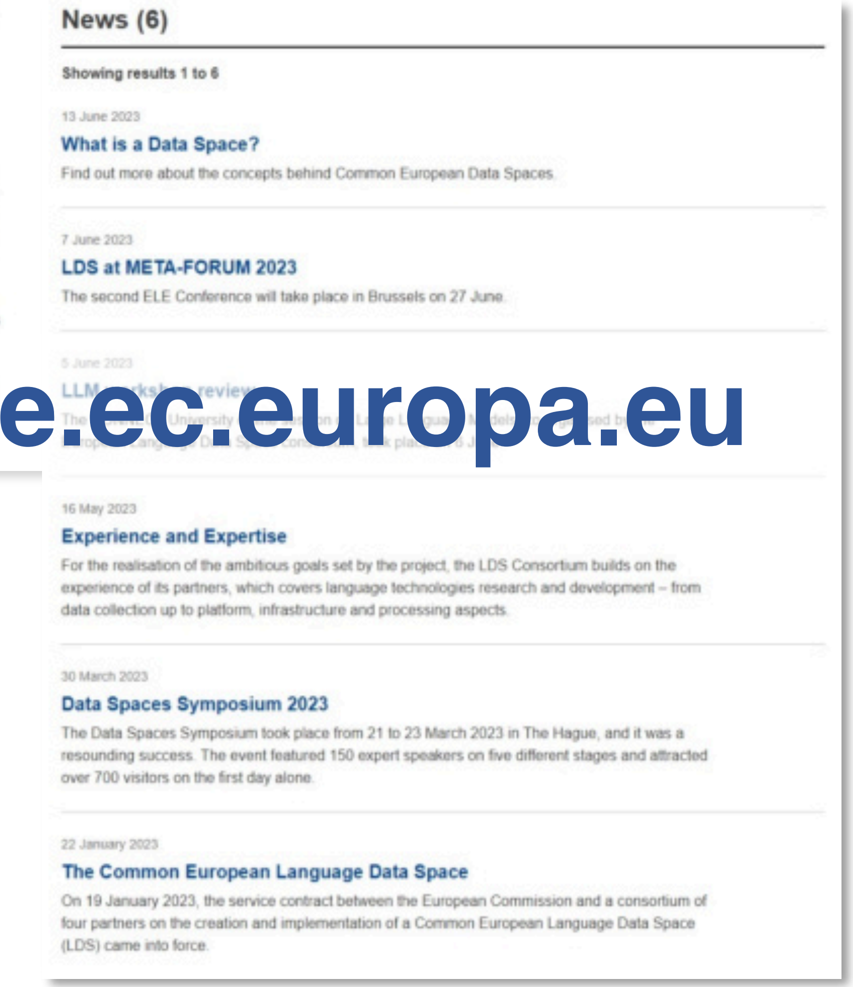
On 19 January 2023, the service contract between the European Commission and a consortium of four partners on the creation and implementation of a Common European Language Data Space came into force.

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



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

European Language Equality

EUROPEAN LANGUAGE EQUALITY

Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

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

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu