



EUROPEAN LANGUAGE DATA SPACE



LDS Country Workshop Luxembourg

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

19-06-2024 LDS Country Workshop Luxembourg
<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, except English
- A concerted effort for the collection of enormous amounts of language data for all European languages is very much needed
- The global NLP/LT/Gen-AI market is in the hundreds of billions of US-\$ already and expected to grow to 439.85B US-\$ by 2030 – no significant players from Europe
- Already now billions and billions are made but ...

BUSINESS

ChatGPT Shows Just How Far Europe Lags in Tech

Analysis by Lionel Laurent | Bloomberg

February 21, 2023 at 2:12 a.m. EST



Comment 1



Gift Article



Share

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.

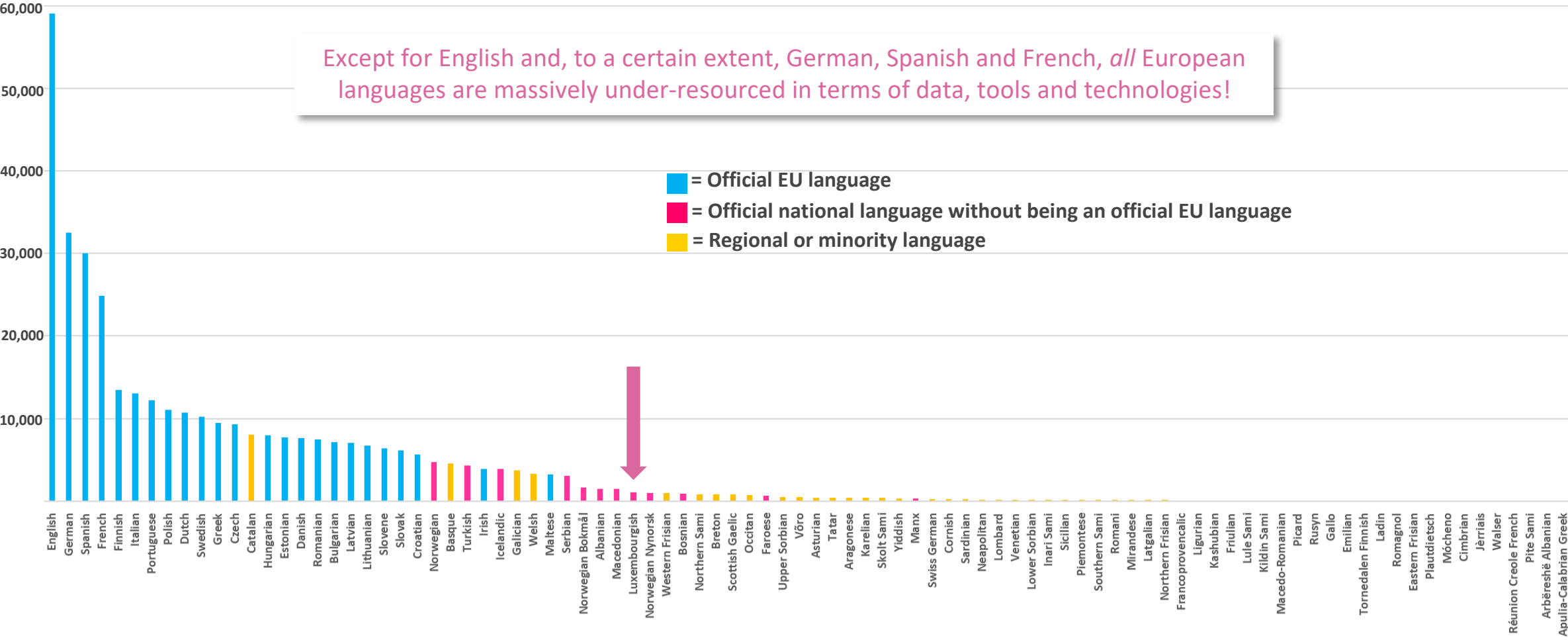
European Initiatives

- 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;
 - HPC facilities;
 - Speed of the big tech players in the US and Asia vs. speed of Europe

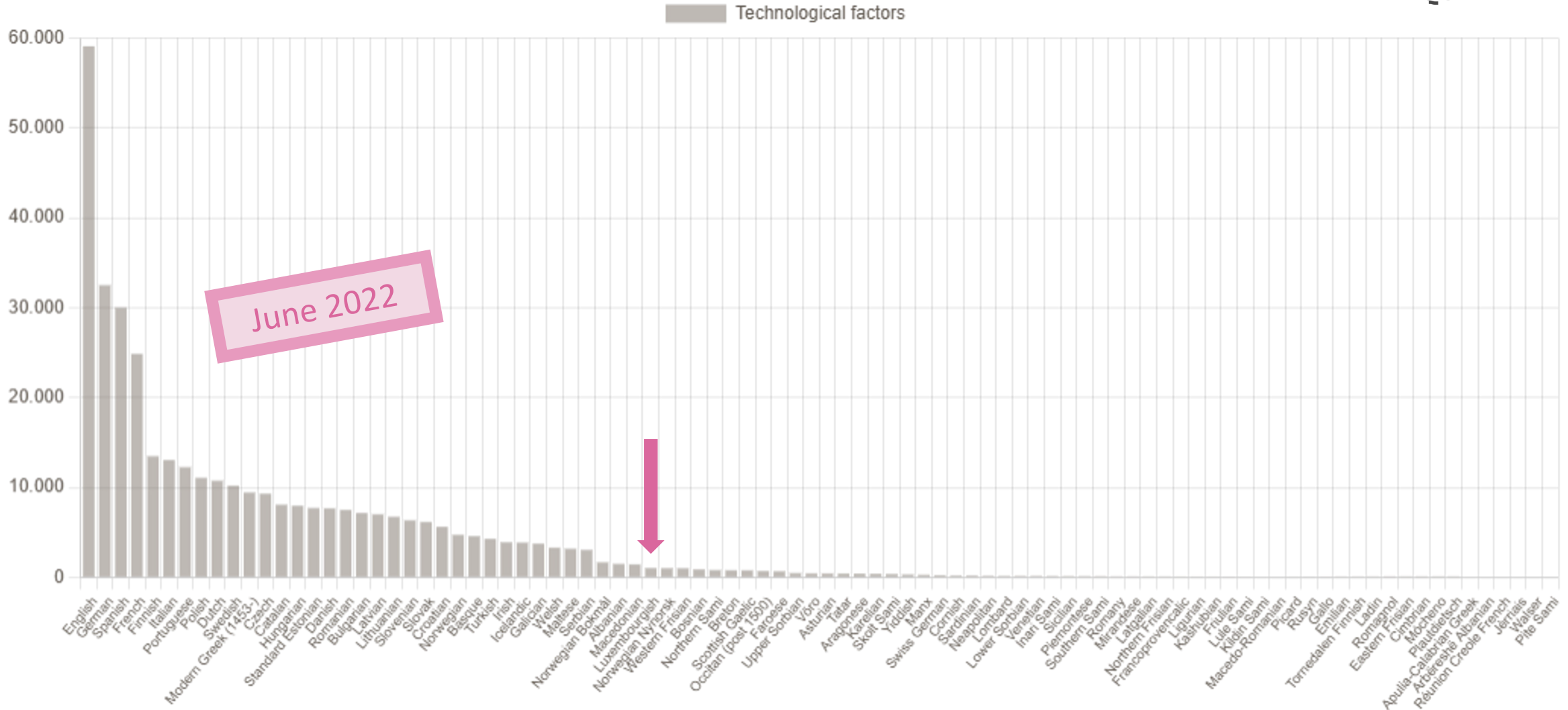
Digital Language Equality Metric: Technological Scores

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!

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



DLE Metric: 2022 vs. 2023 (1/3)



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 14 official EU data space projects with a strong focus on industry

Common European Language Data Space



**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 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
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

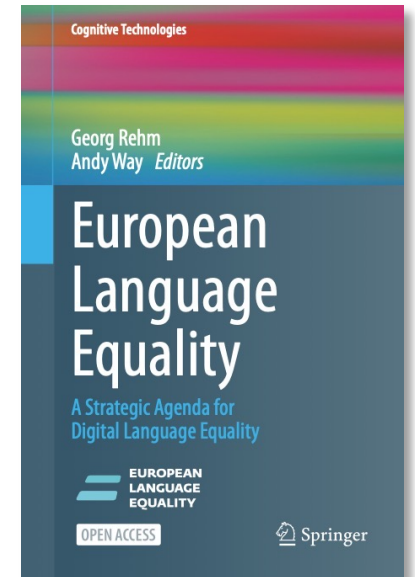
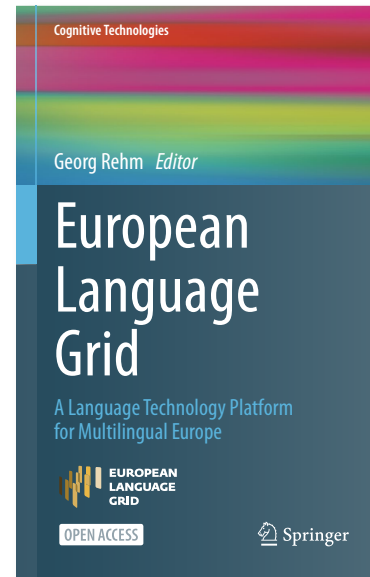
Previous Projects and Initiatives

- The four core partners – DFKI, ILSP, ELDA, TILDE – 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)

META  **NET**



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



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: Media Data Space	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

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

7 observing Member States: Austria, Belgium, Estonia, Malta, Portugal, Romania and Slovakia.

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](#) (returning user) and go on to describe

ELRA ASSOCIATION OF EUROPEAN LANGUAGE RESOURCES

1096 Language Resources (Page 1 of 55)

2006 CoNLL Shared Task - Arabic & Czech

& Czech consists of dependency treebanks used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The Conference on Learning (CoNLL) is accompanied every year by a shared task intended to promote natural lan...

License	Non-Commercial/Yes - Non-Standard License Terms	Commercial
CC BY	academic	commercial
CC BY-NC	academic	commercial

- Ten Languages

Japanese / Portuguese / Slovenian / Spanish / Catalan / Swedish / Turkish

EUROPEAN LANGUAGE GRID

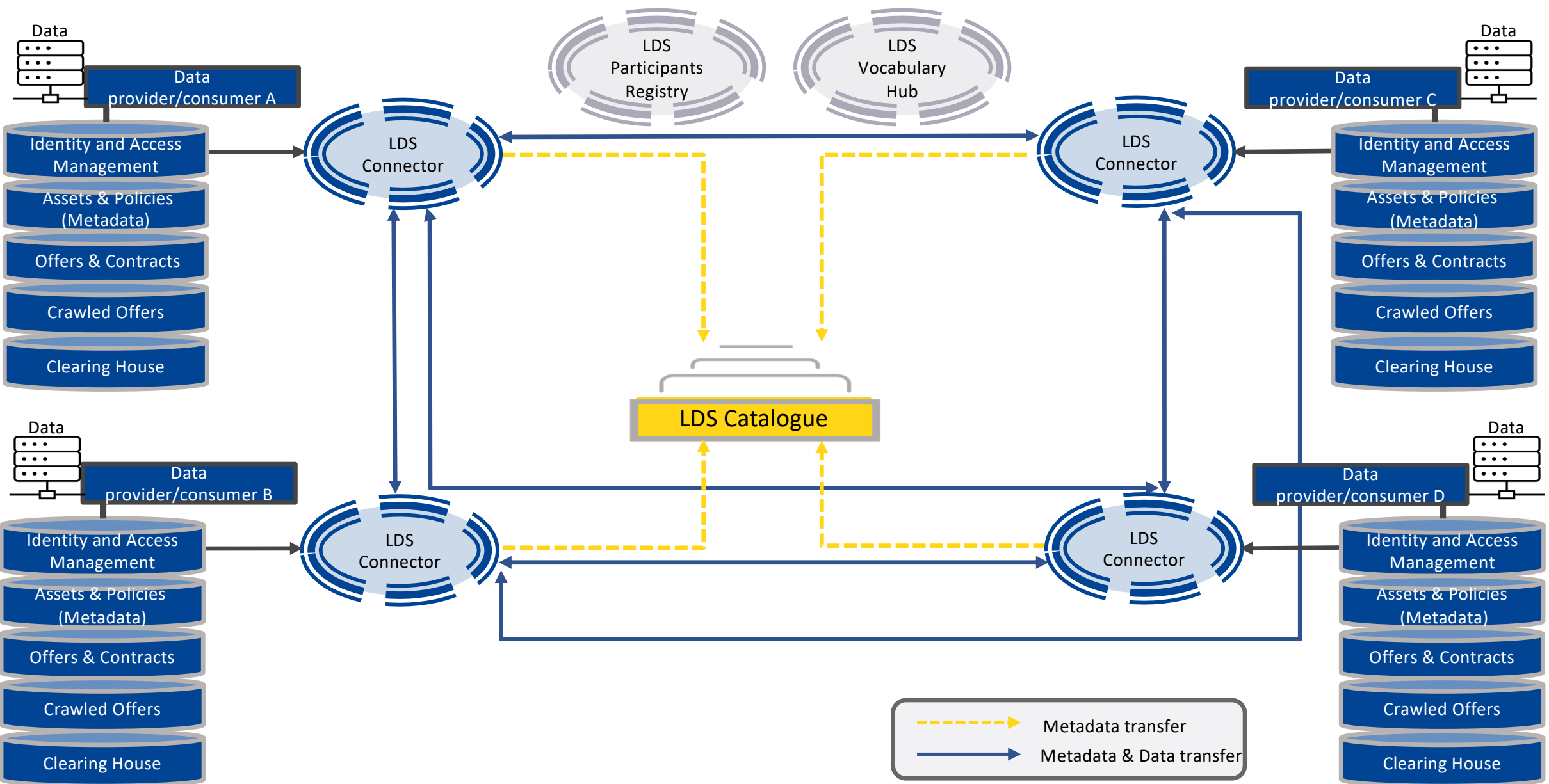
Language Technologies

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.

Search the catalogue

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



Current 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.

Assets
15

Create A New
Asset

View My
Assets

Policies
16

Define A New
Policy

View My
Policies

Offers
6

Create Offers

View Offers

Create new data asset

Create a new asset

Select language (optional)
Language
ENGLISH

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

Privacy properties
anonymization or sensitive data details.
Privacy

Language
A language of the resource.
Language

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

IprHolder properties
iprHolder or creator details...
IPR holder

Documentation
related documentation
is documented by

Temporal properties
time constraints
Temporal Coverage

Identifiers
identifier details
Identifiers

Distribution
media type, format, ...
distribution

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

SAVE ASSET

Current LDS
Prototype

Create and adjust policy

POLICY CLASS

✓ Interval-restricted Data Usage Policy Class
allows data usage for a specified time period

Purpose-restricted Data Usage Policy Class
allows data usage for a specific declared purpose

pending

Connector-restricted Data Usage Policy Class
allows data usage for a specific connector

pending

Perpetual Data Sale (Payment once) Policy Class
allows data usage after payment is completed (for datasets requiring once-off payment)

pending

Location Restriction of the participant for Data Usage Policy Class
restricts data usage to participants in a specific location

pending

Attribution Data Policy Class
allows distribution of data with mandatory attribution

pending

Share Alike derivatives Policy Class
allows distribution of derivatives with a compatible

pending

Attach Policy When Distribute to a third-party Policy Class
allows distribution of data to a third-party with the specified attached policy

pending

Derivatives not allowed Policy Class
distribution of derivatives is not allowed

pending

Current LDS
Prototype

Create and publish offer

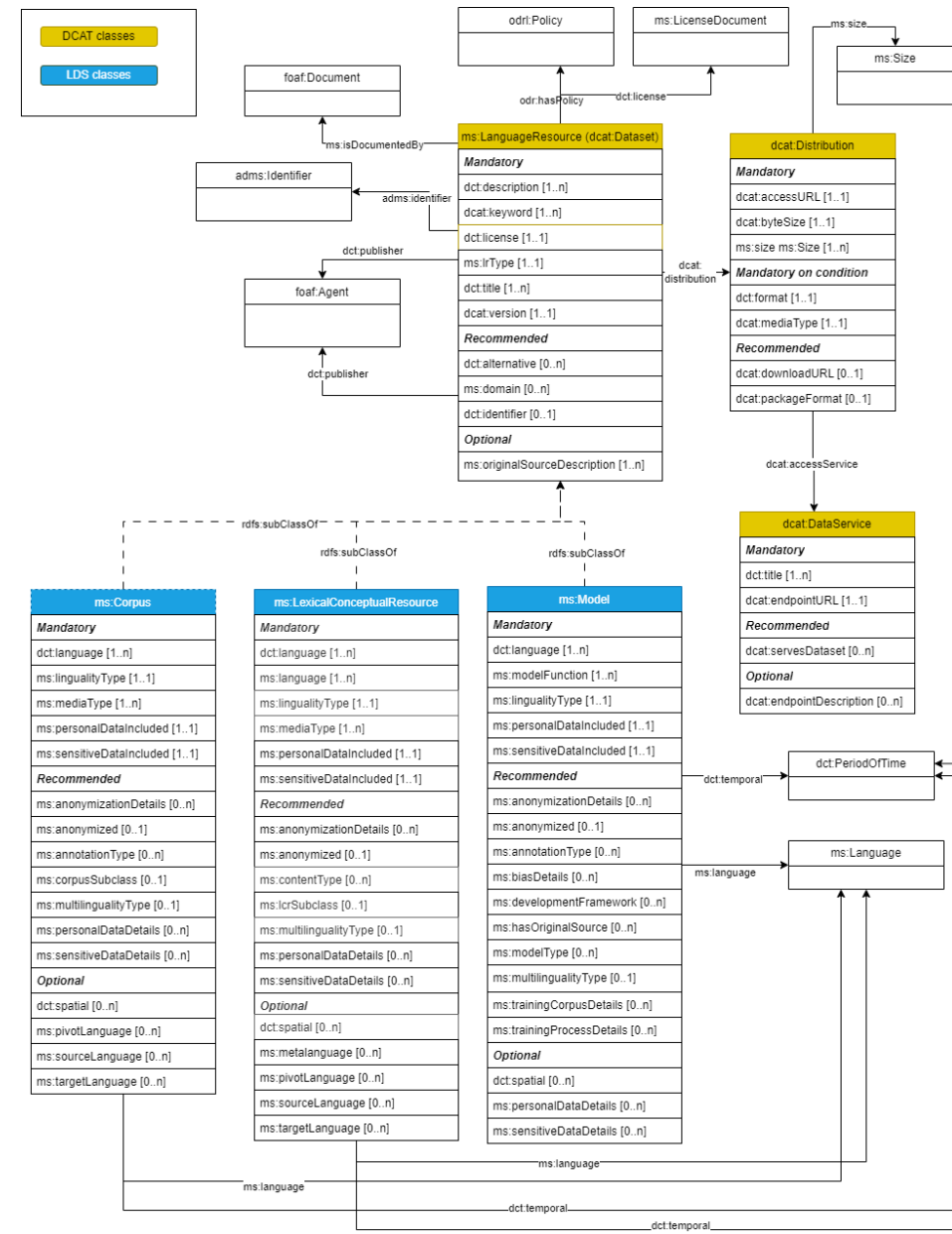
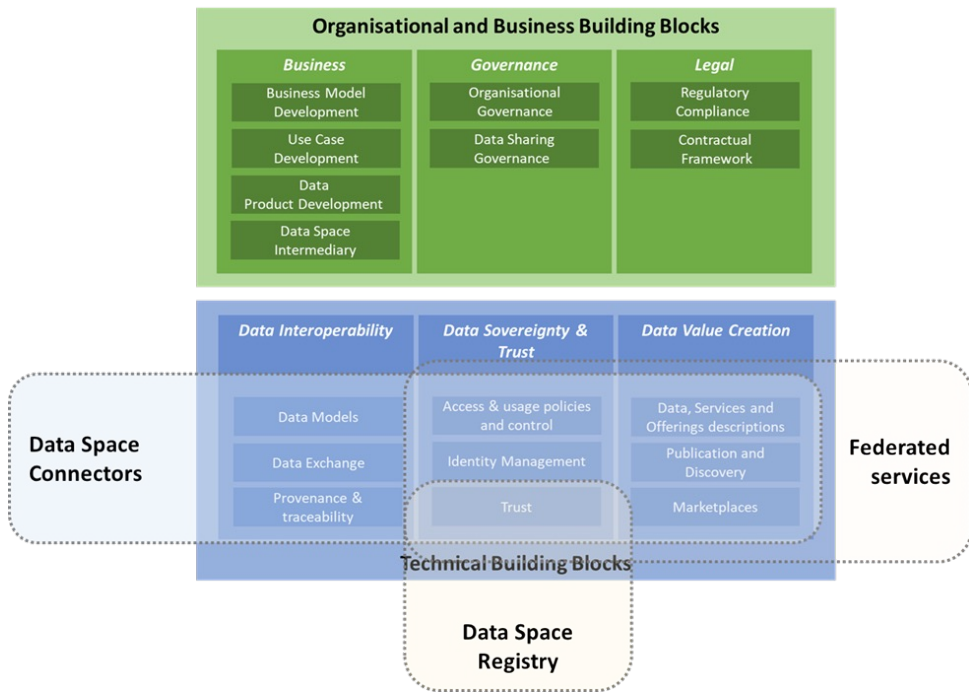
Current LDS
Prototype

Progress: 1 Select Asset — 2 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 ▼



Build on Existing Solutions

- Following DSSC (see above)
- Eclipse Data Space Components (EDC)
- DCAT-AP, Language DCAT-AP (see right), ODRL
- Mappers from existing platforms

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

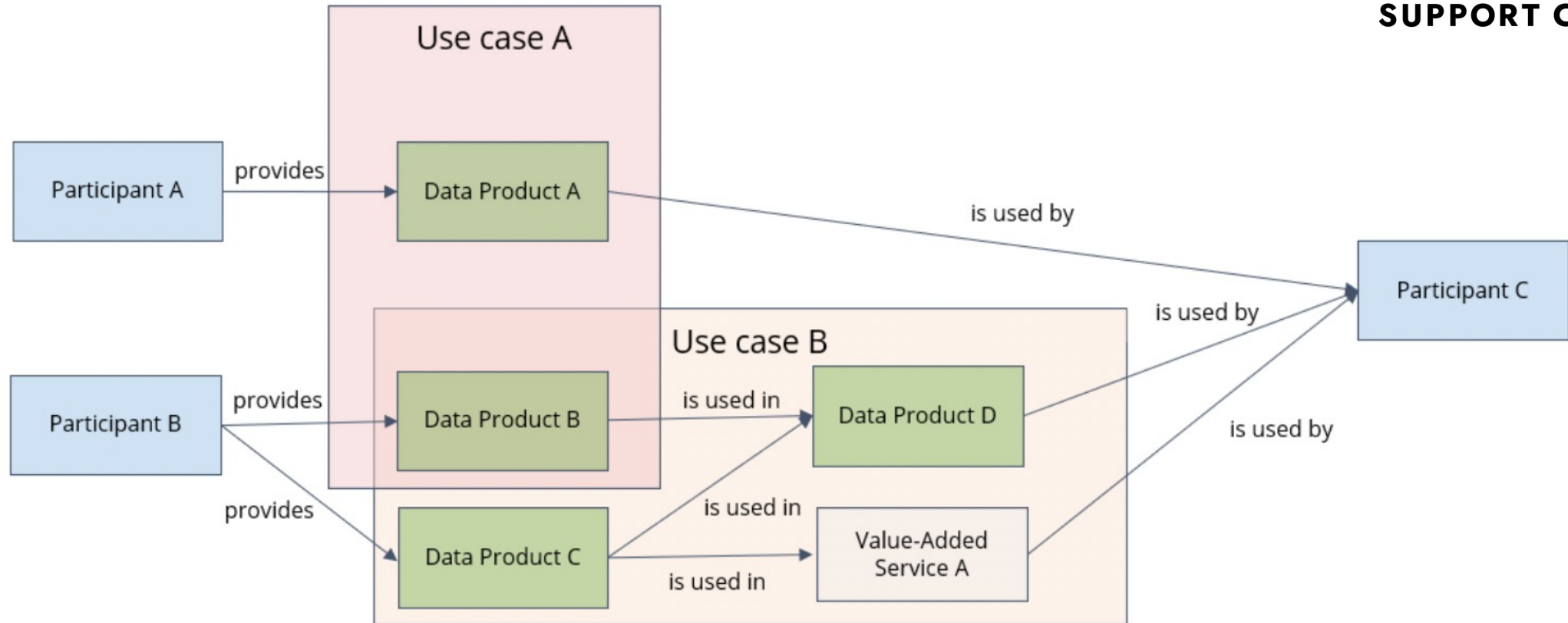
- Buy or access data products to develop better services (including LLMs)
 - Multilingual data
 - Multimodal data
 - Domain-specific data
 - All European languages
 - Easy discoverability and access
- Limited effort: keep existing infrastructure
- Legal compliance by design
 - Compliance with EU regulation and standards
 - Transparency: emphasis on data provenance
- Find new customers for services and products

Data Products

DSSC Blueprint 1.0 – Data Product Development



**DATA SPACES
SUPPORT CENTRE**



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.)
- *A language resource corresponds to a data product*
- Of utmost importance now: large amounts of language data to pre-train large language models
- LRs were typically available for research purposes free of charge
- LRs were sometimes available for commercial use for a certain fee
- Many unique LRs developed by European research organisations were licensed by various European but also large US tech companies, e.g., for online NLP services (Machine Translation)

REUTERS World Business Markets Sustainability Legal Breakthroughs Technology Investigations More My View Sign In Register

Exclusive news, data and analytics for financial market professionals

Exclusive: Reddit in AI content licensing deal with Google


By Anna Tong, Echo Wang and Martin Coulter
February 22, 2024 5:10 AM GMT+1 Updated 2 months ago



Reddit logo is seen in this illustration taken November 2, 2022. REUTERS/Cweb/Rover/illustration/Fix Photo/Duckart/Anastasia Stets 13

ARTIFICIAL INTELLIGENCE / TECH

OpenAI's news publisher deals reportedly top out at \$5 million a year



The ChatGPT company has been trying to get more news organizations to sign licensing deals to train AI models.

By Emily Davis, a reporter who covers AI. Prior to joining The Verge, she covered the intersection between technology, finance, and the economy.
Apr 4, 2024, 8:39 PM GMT+1

As news publishers ink deals with AI companies to train their models with new stories, the price businesses like OpenAI are willing to pay for copyrighted information is coming to light.

The *Information* reports that OpenAI offers between \$1 million and \$5 million a year to license copyrighted news articles to train its AI models. That's one of the first indications of how much AI companies

Bloomberg Sign In Subscribe

Live TV Markets Economics Industries Tech Politics Businessweek Opinion More Europe Edition

Technology AI

OpenAI In Talks With CNN, Fox and Time to License Content

- Startup has said it's in discussions with dozens of publishers
- Negotiations come as OpenAI faces New York Times lawsuit

By Shirin Ghaffary, Graham Starr, and Brody Ford
10 January 2024 at 23:52 CET
Updated on 11 January 2024 at 17:43 CET

OpenAI is in talks with CNN, Fox Corp. and Time to license their work, according to people familiar with the matter, in a growing effort to secure access to news content to build out its artificial intelligence products while facing allegations it's ripping off copyrighted materials.

The startup behind ChatGPT, a tool that lets users quickly crank out text, code and other content with simple prompts, is seeking to cut


Have a confidential tip for our reporters? Get in Touch

Create your account to continue reading.

OpenAI Research API ChatGPT Safety Company Search Log in Try ChatGPT

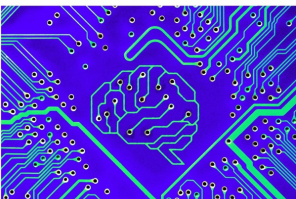
Partnership with Axel Springer to deepen beneficial use of AI in journalism

Axel Springer is the first publishing house globally to partner with us on a deeper integration of journalism in AI technologies.



ARTIFICIAL INTELLIGENCE / TECH / GOOGLE

OpenAI transcribed over a million hours of YouTube videos to train GPT-4



A New York Times report details the ways big players in AI have tried to expand their data access.

By Will Davis, a weekend editor who covers the latest in tech and entertainment. He has written news, reviews, and more as a tech journalist since 2020.
Apr 6, 2024, 10:29 PM GMT+1

Earlier this week, *The Wall Street Journal* reported that AI companies were running into a wall when it comes to gathering high-quality training data. Today, *The New York Times* detailed some of the ways companies have dealt with this. Unsurprisingly, it involves doing things that fall into the hazy gray area of AI copyright law.

Le Monde NEWS INTERNATIONAL VIDEOS ENVIRONMENT FRANCE OPINION FRENCH DELIGHTS Subscribe

EDITORIAL

Louis Dreyfus
Chief Executive Officer of Le Monde
Jérôme Fenoglio
Director of Le Monde

Le Monde and Open AI sign partnership agreement on artificial intelligence

This multi-year agreement, the first between a French media organization and a major AI player, will enable OpenAI to draw on our newspaper's corpus to establish and enhance the reliability of the answers of its ChatGPT tool, in return for a significant source of additional revenue.

Published on March 13, 2024, at 6:31 pm (Paris), updated on March 13, 2024, at 9:05 pm 0 min. Lire en français

A part of its discussions with major players in the field of artificial intelligence, *Le Monde* has just signed a multi-year agreement with OpenAI, the company known for its ChatGPT tool. This agreement is historic as it is the first signed between a French media organization and a major player in this nascent industry. It covers both the training of artificial intelligence models developed by the American company and answer engine services such as ChatGPT. It will benefit users of this tool by improving its relevance thanks to recent, authoritative content on a wide range of current topics, while explicitly highlighting our news organization's contribution to OpenAI's services.





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](#)



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 for data licenses
- This is LDS's baseline!
- 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

Next Steps

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 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

LDS User Group

• Meetings and Conferences

- Inaugural meeting in March 2024
- Second meeting in June 2024
- Next meeting in September 2024
- LDS Launch Conference: October 2024

• Communication

- Established a mailing list for the LDS user group
- The LDS user group will grow – new members will be added to the mailing list

• **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 (foreseen for Q3 2024); 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.

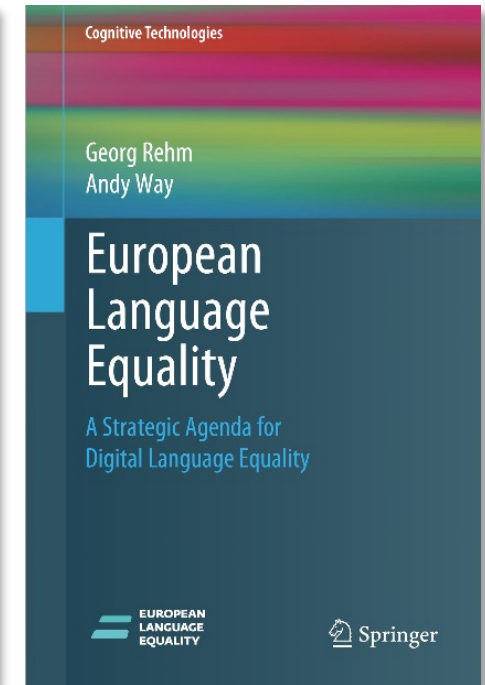
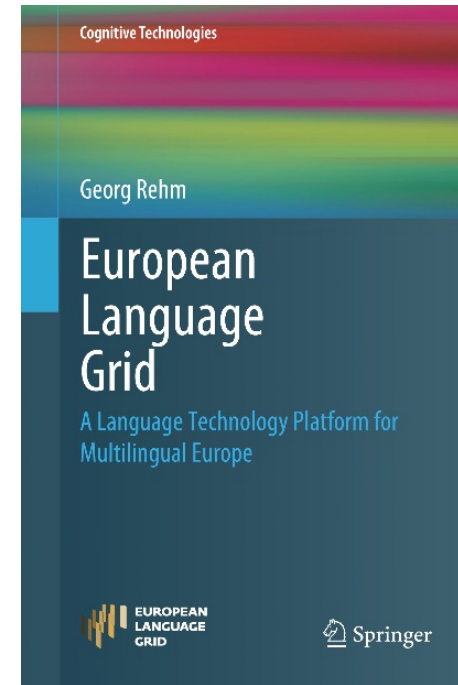
[Click to join](#)





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) – LDS Coordinator
georg.rehm@dfki.de

19-06-2024 LDS Country Workshop Luxembourg
<https://language-data-space.ec.europa.eu>