EUROPEAN LANGUAGE DATA SPACE

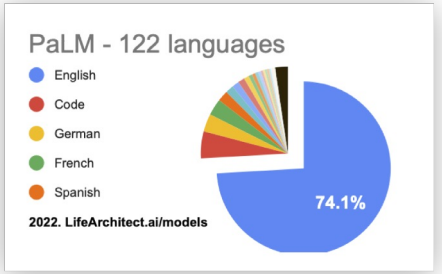
LDS Country Workshop Slovakia

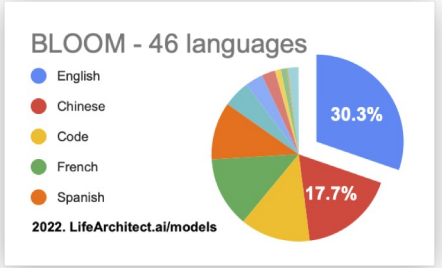
Dr. Simon Ostermann (DFKI GmbH, Germany) – LDS Project Manager and Site Manager Saarbrücken
simon.ostermann@dfki.de

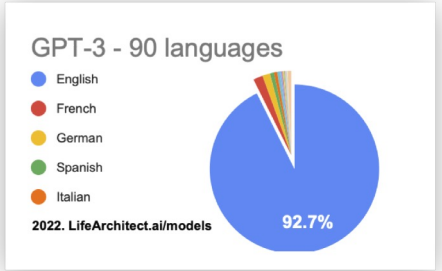
07-11-2024 LDS Country Workshop Slovakia
https://language-data-space.ec.europa.eu

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 …







The Washington Post
Democracy Dies in Darkness

Sign in

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

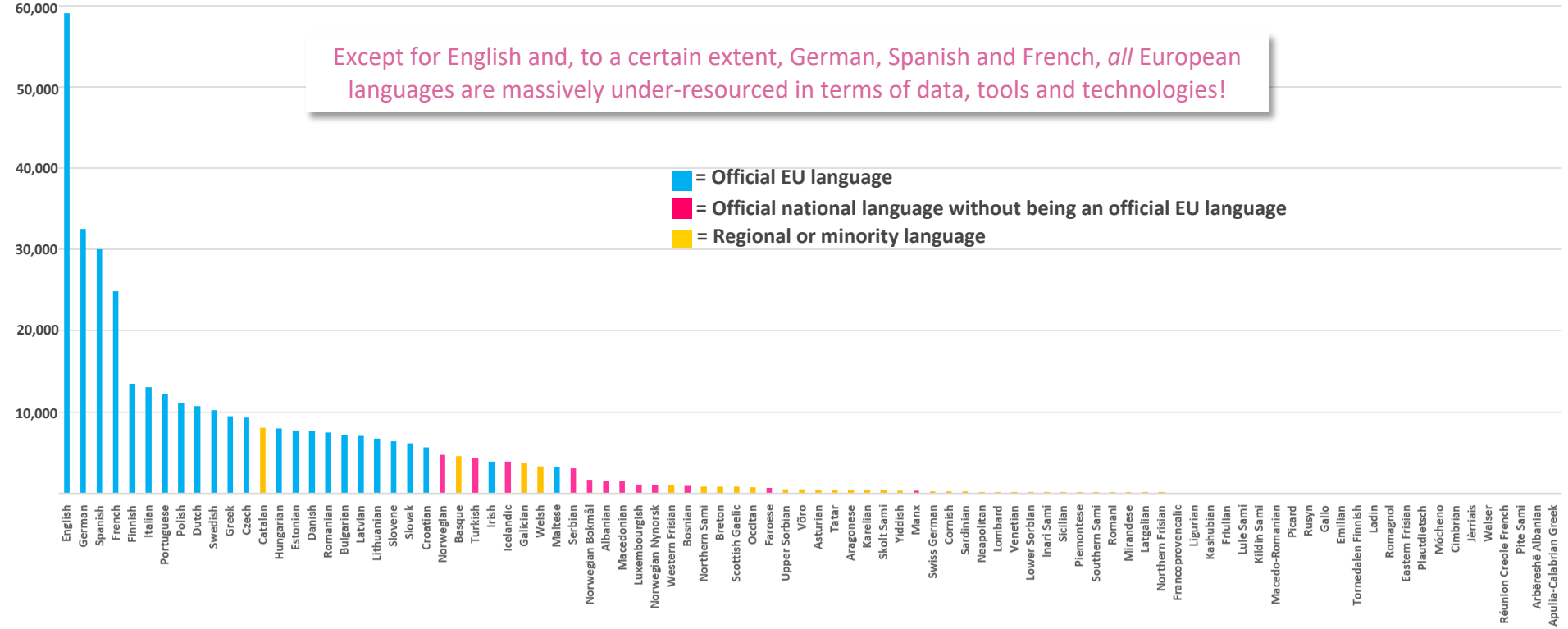
Europe is where ChatGPT gets regulated, not invented. That's something to regret. As unhinged as the initial results of the artificial-intelligence arms race may be, they're also another reminder of how far the European Union lags behind the US and China when it comes to tech.

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

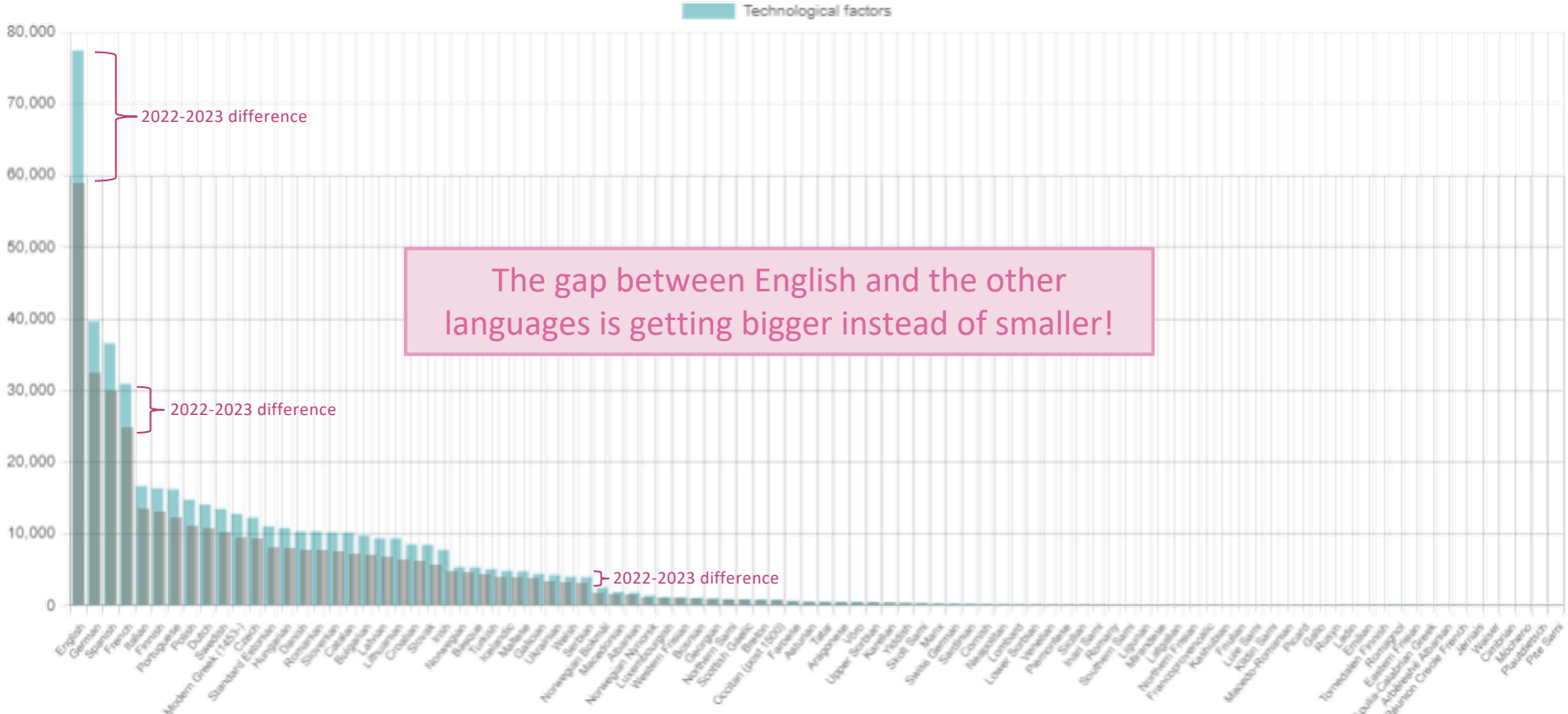
Digital Language Equality Metric: Technological Scores

EUROPEAN LANGUAGE EQUALITY



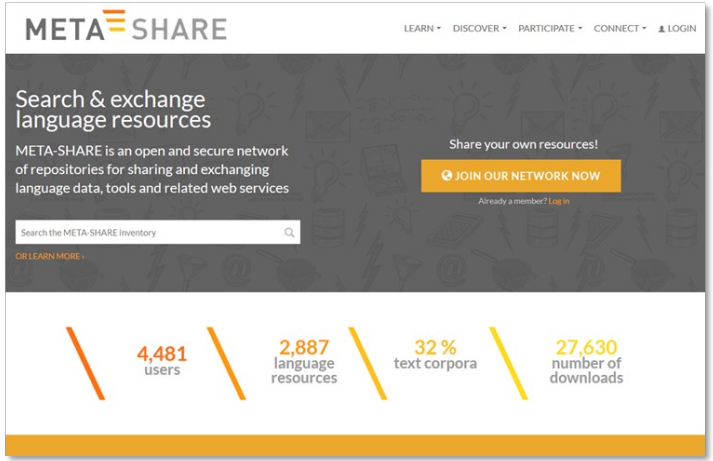
DLE Metric: 2022 vs. 2023

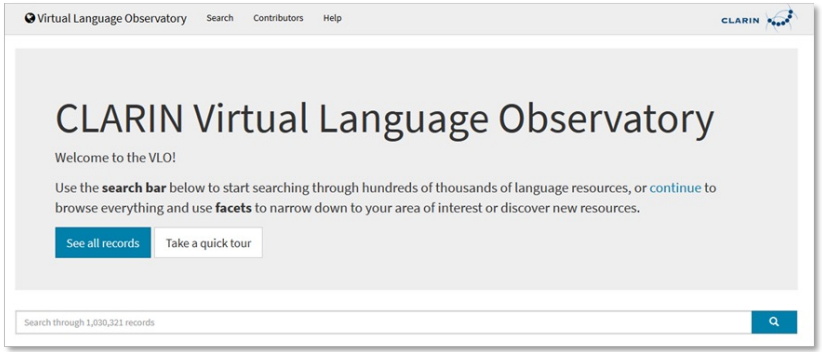
EUROPEAN LANGUAGE EQUALITY

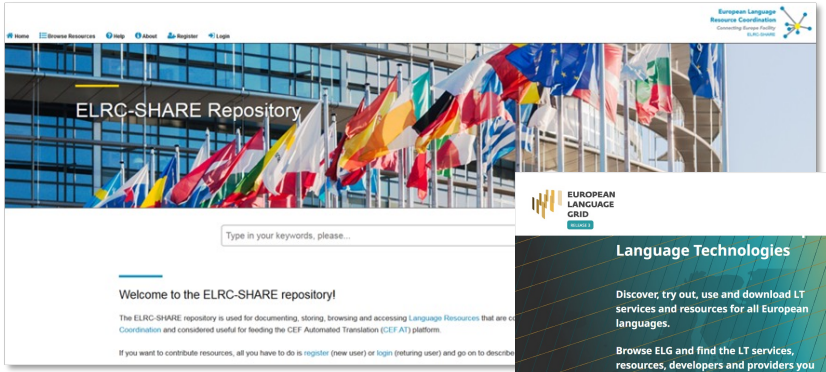


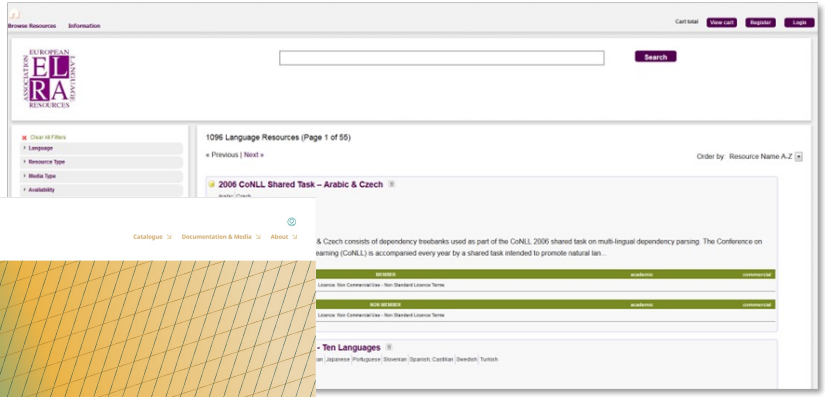
EU Data Strategy

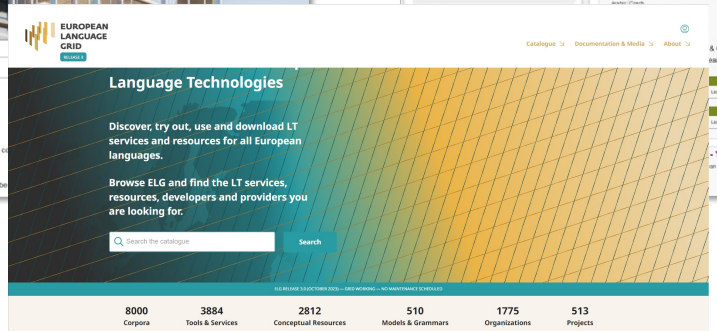
Long History of Language Data Sharing







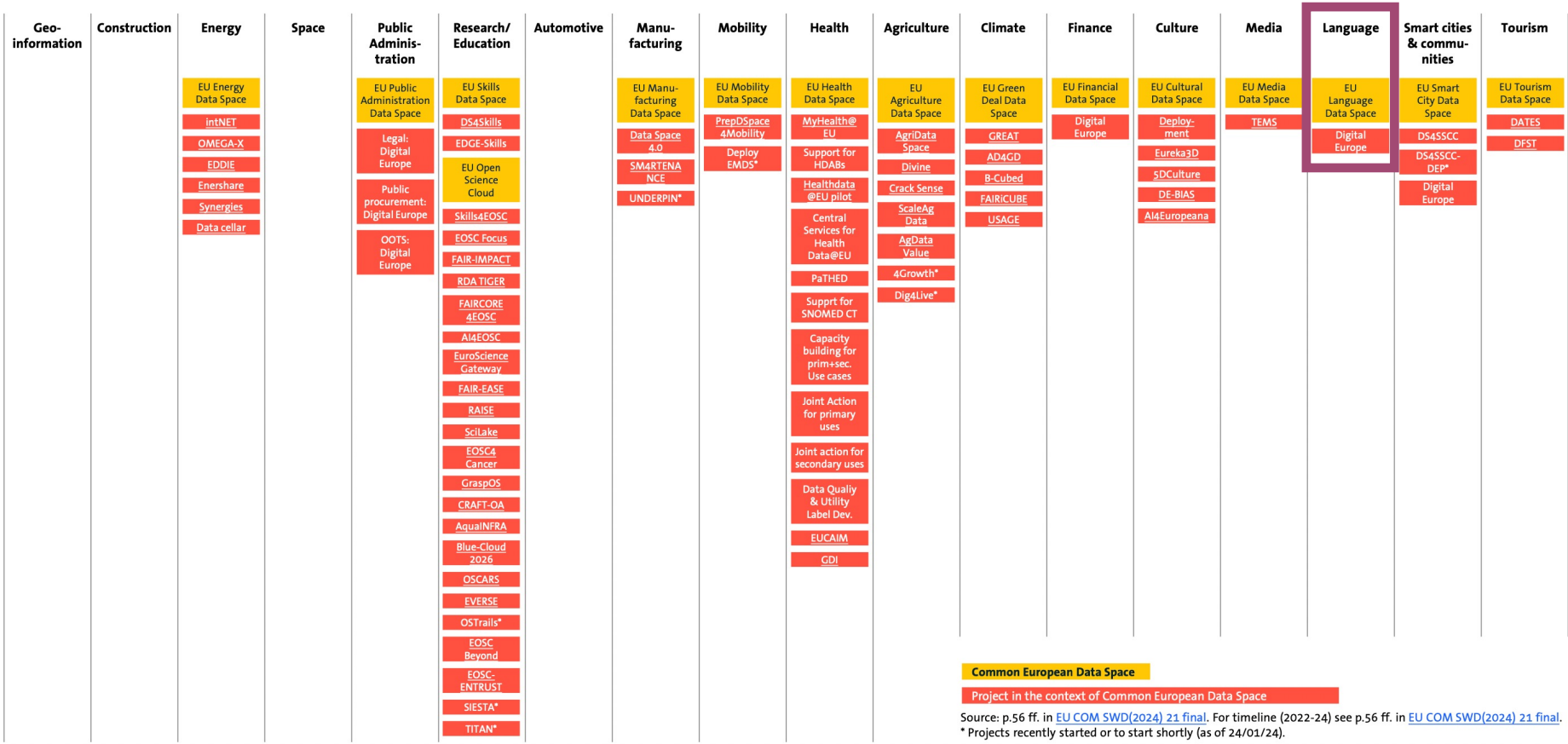




EU Data Strategy & Data Spaces

- 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 15 official EU data space projects with a strong focus on industry

Data initiatives – EU-level



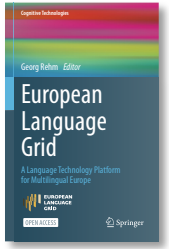
Source: p.56 ff. in EU COM SWD(2024) 21 final. For timeline (2022-24) see p.56 ff. in EU COM SWD(2024) 21 final.
* Projects recently started or to start shortly (as of 24/01/24).

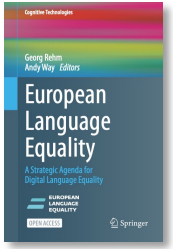
Source: Bitkom “Data Economy Landscape”, April 2024, Version 1.1

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
- 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

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

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)

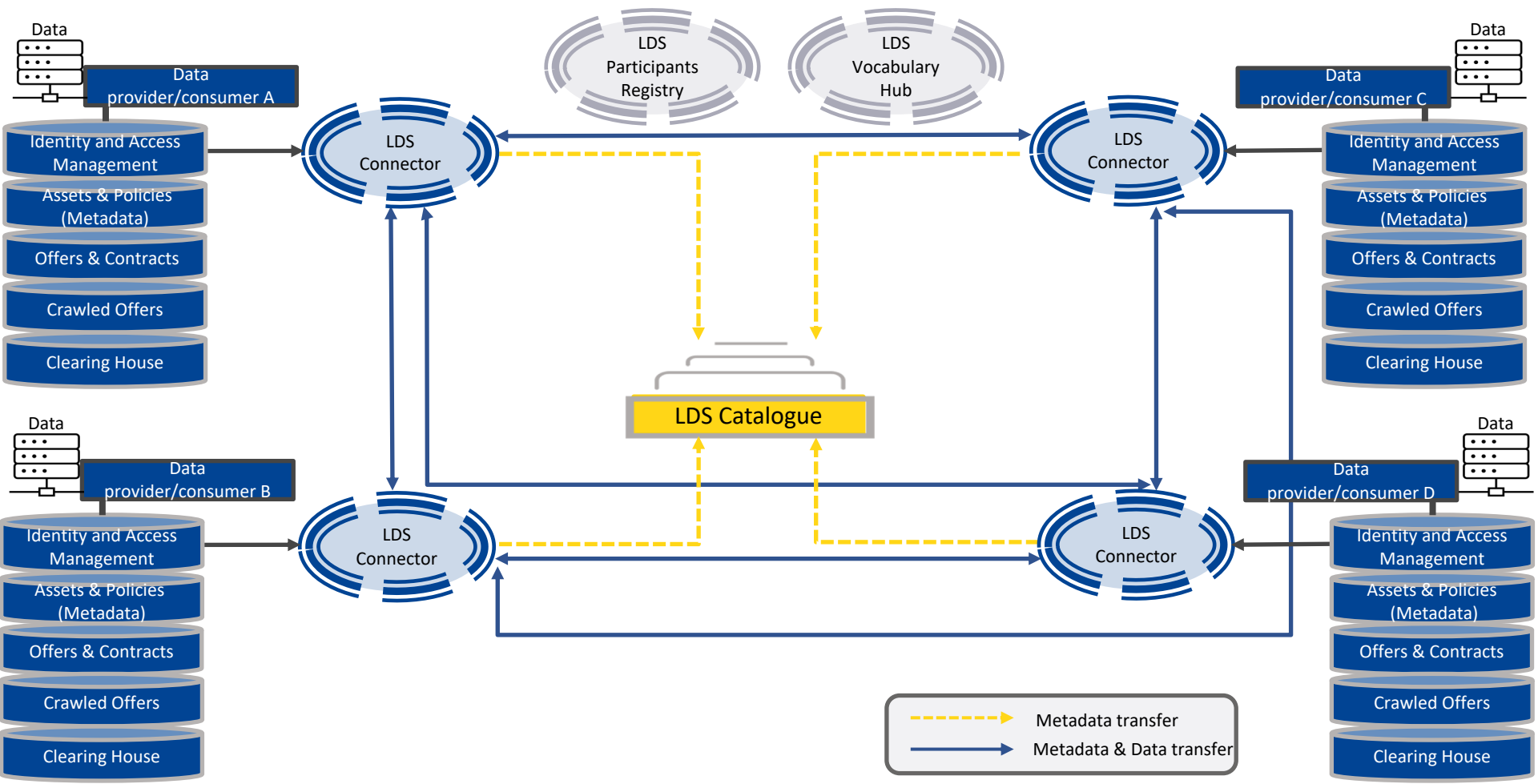
- European Digital Infrastructure Consortium (EDIC): a new legal entity type in the EU
- The first couple of EDICs are currently under development including the ALT-EDIC
- Coordinated by the French Ministry of Culture
- Close collaboration between: ALT-EDIC Working Group, EC, LDS
- ALT-EDIC action plan 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

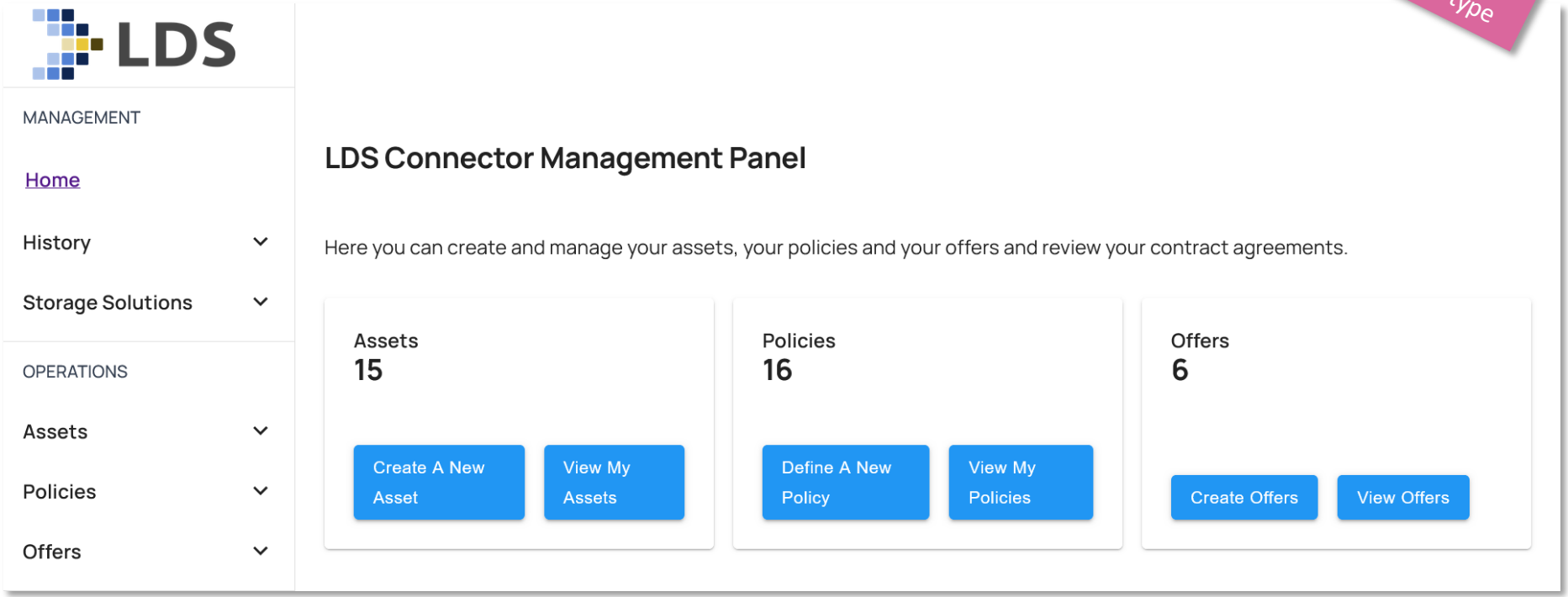
17 Members States: Bulgaria, Croatia, Czechia, Denmark, Finland, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Slovenia and Spain;

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

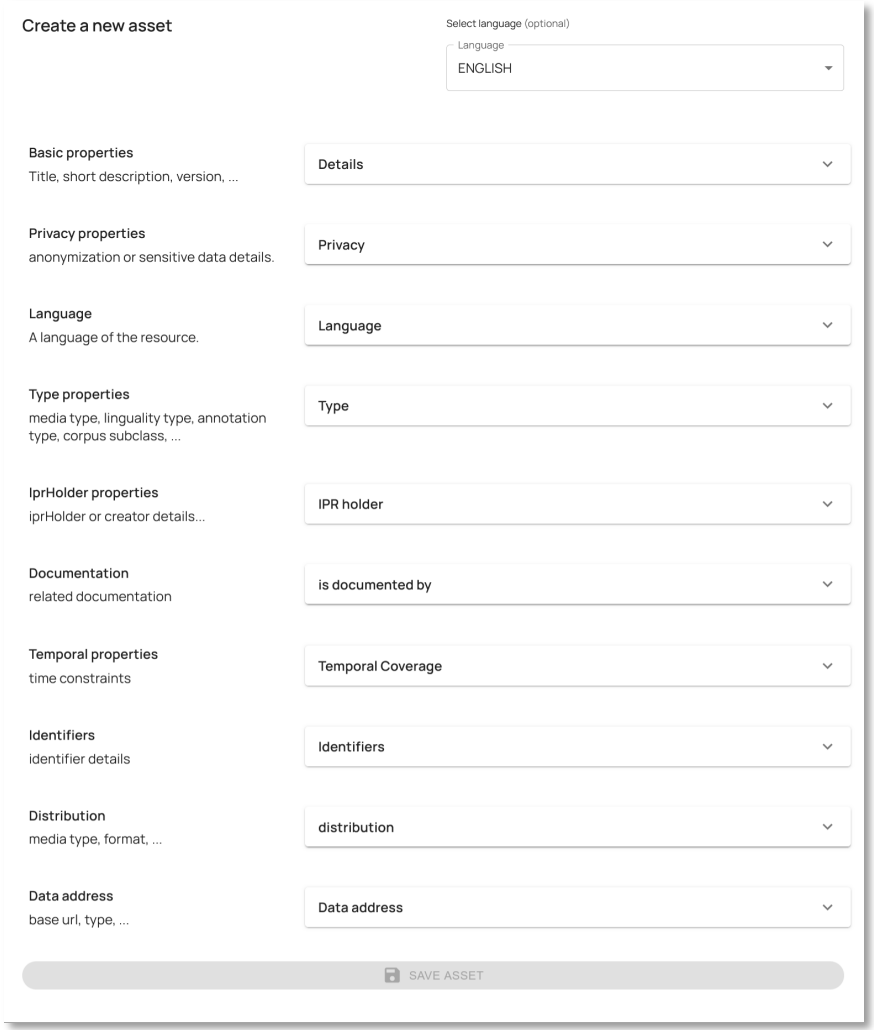
The Common European Language Data Space - Implementation


Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Participants Registry
LDS Vocabulary Hub
LDS Connector
LDS Connector
Data
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Catalogue
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Connector
LDS Connector
Data
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Metadata transfer
Metadata & Data transfer

Early LDS Prototype

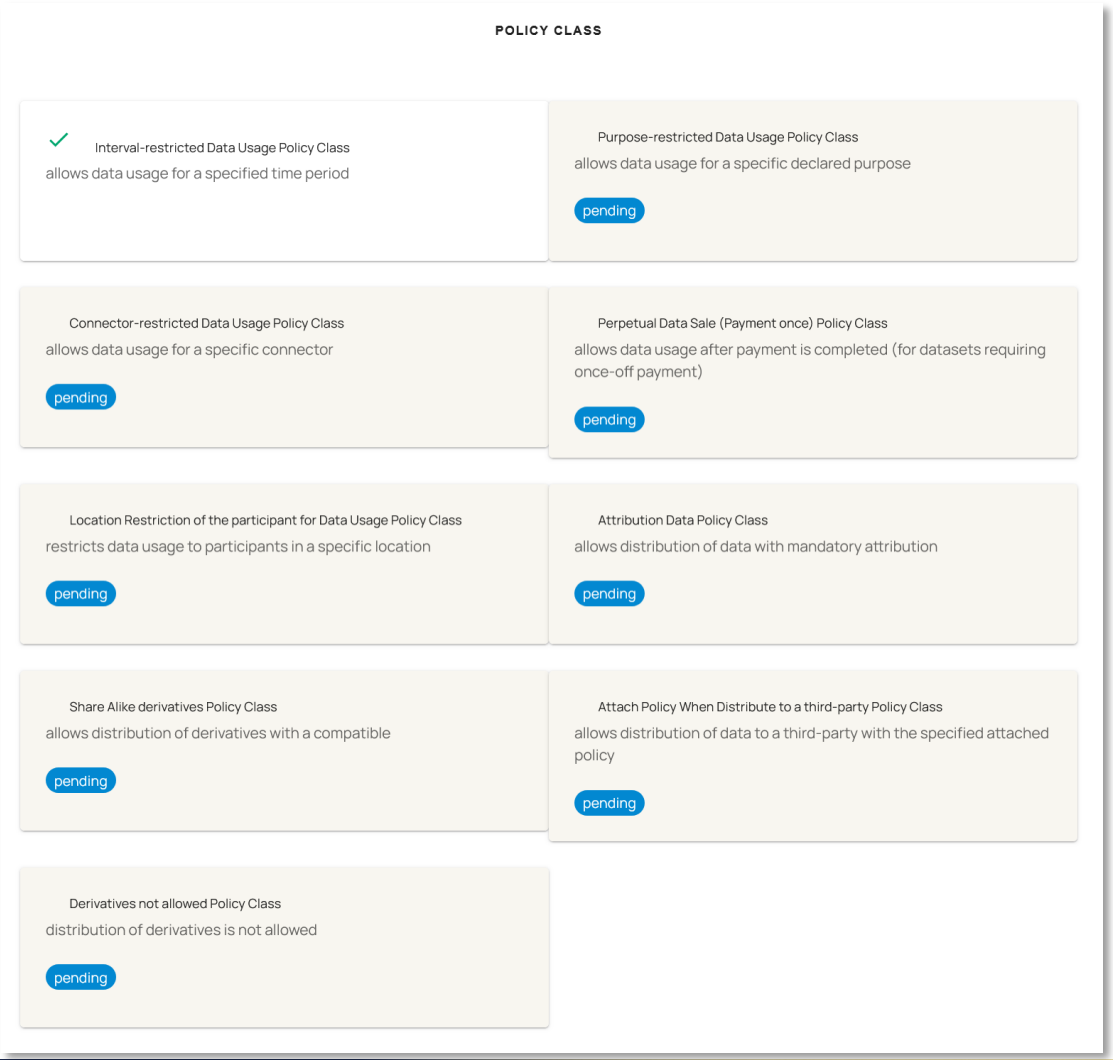


Step 1: Create new data asset



Early LDS Prototype

Step 2: Create and adjust policy

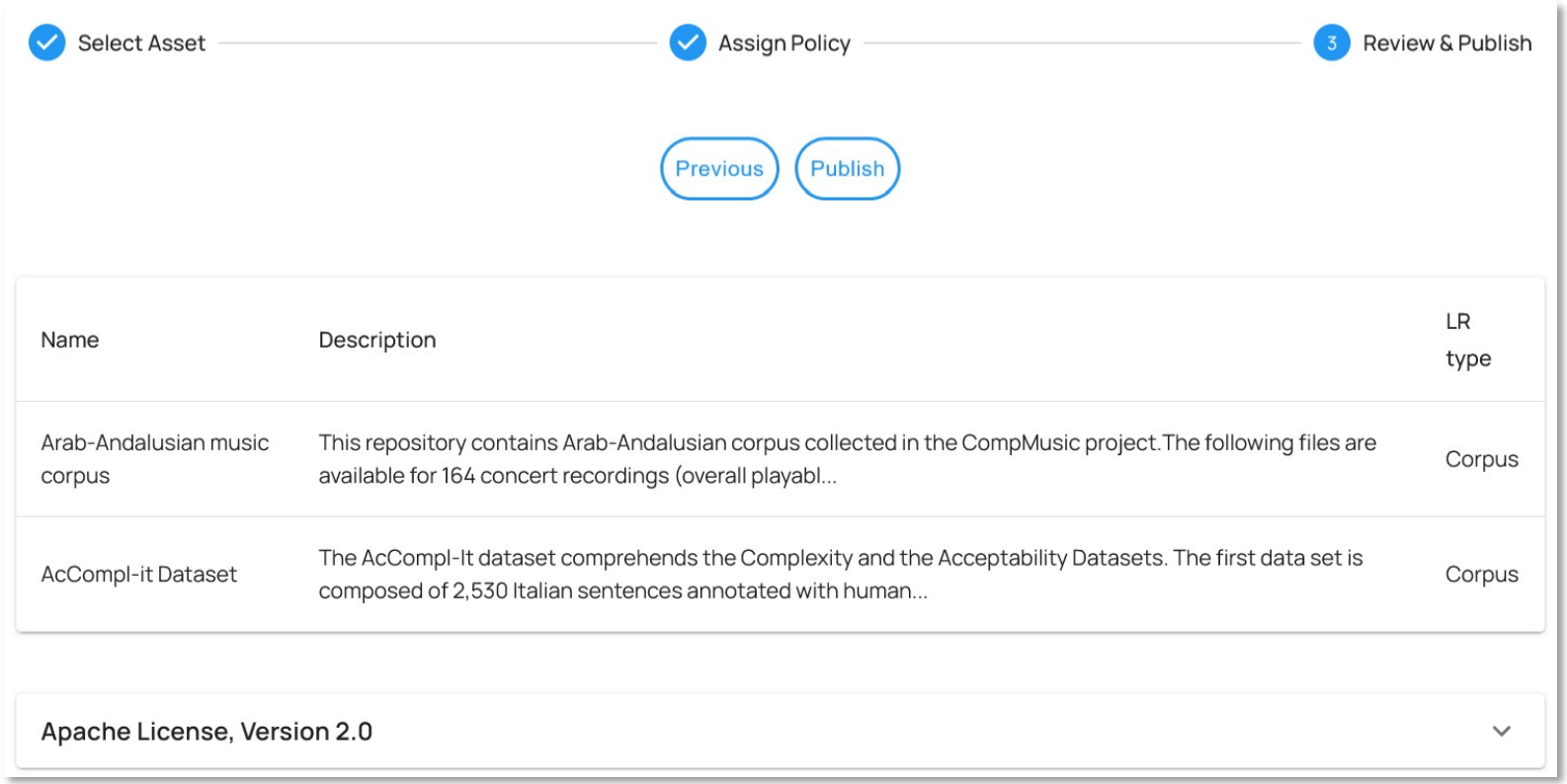


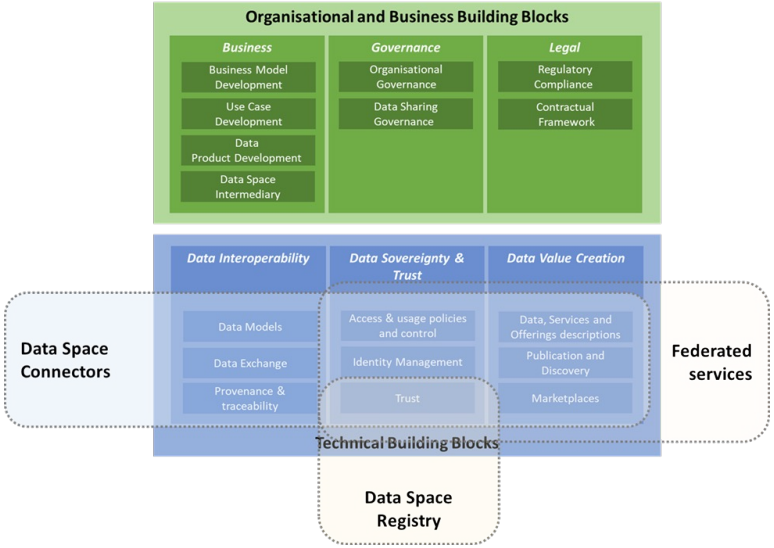
Early LDS Prototype

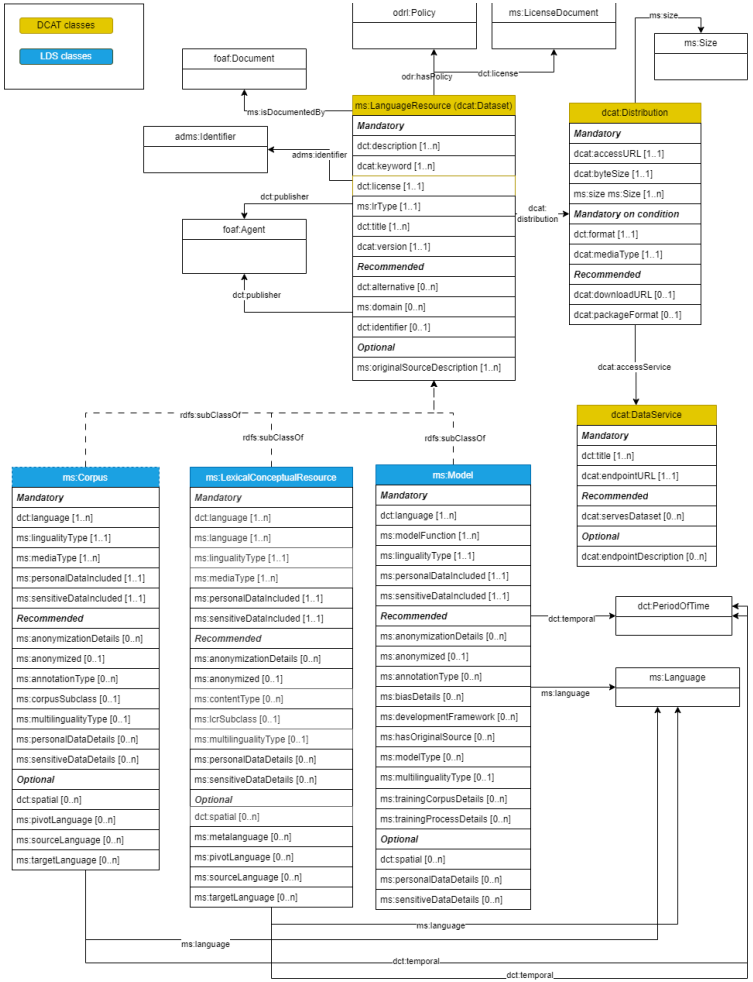
LDS

Step 3: Create and publish offer

Early LDS Prototype







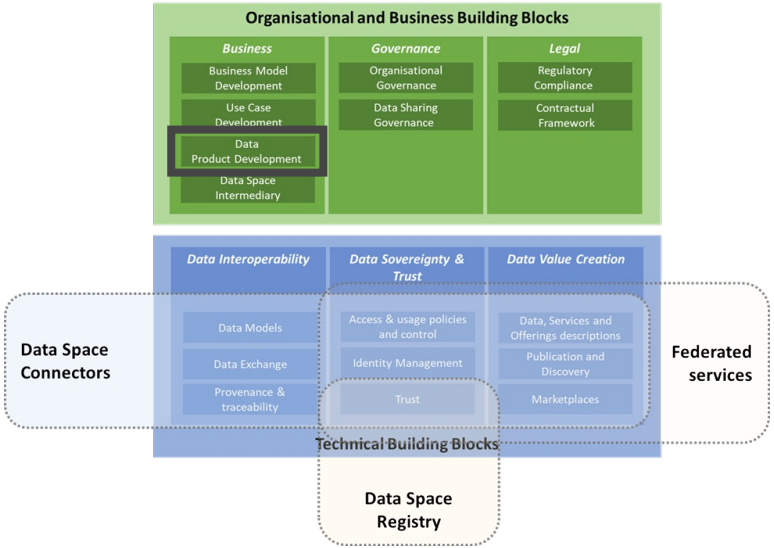
Build on Existing Solutions

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

The Common European Language Data Space – a Value Proposition

DSSC Blueprint 1.0 – Data Product Development

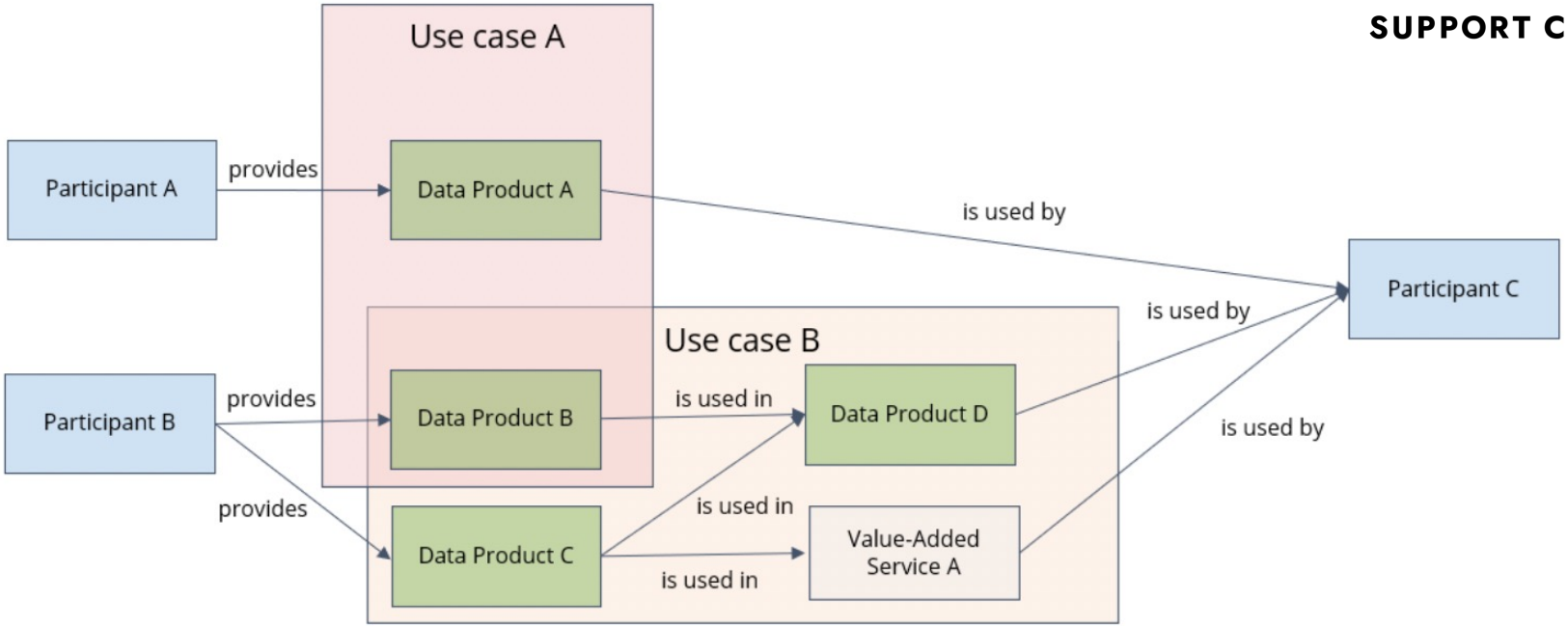
DATA SPACES SUPPORT CENTRE



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.)
- *language resource ≈ 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



DSSC Blueprint 1.0 – Data Product Development

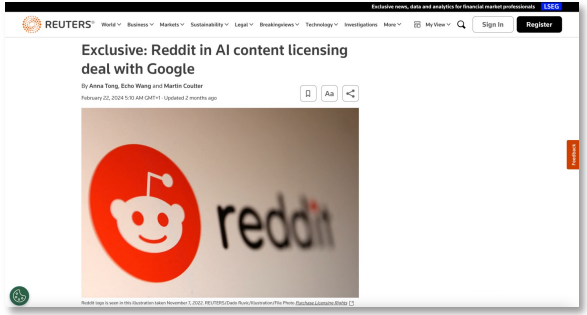
Language Data Space – Value Proposition

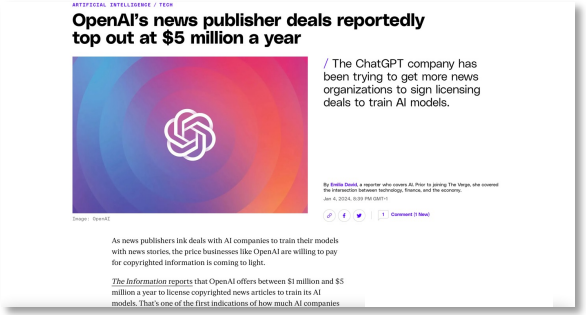
Data Providers

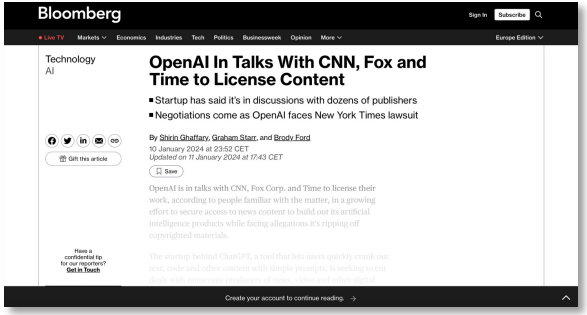
- Additional revenue – LDS as a marketplace
  - Sell data products
  - Find new customers
  - Extend or enrich datasets using AI/NLP services offered in the wider LDS ecosystem
- Legal compliance by design
  - Stay in control over use and access of data
  - Compliance with EU regulation and standards
- 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

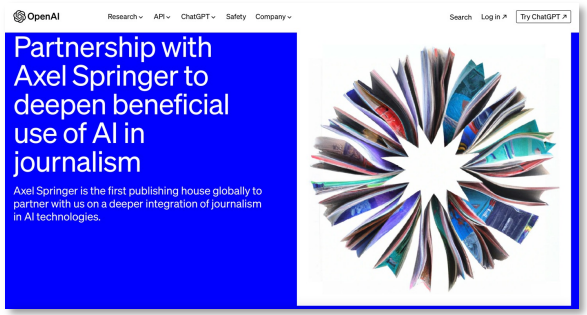
Data Consumers

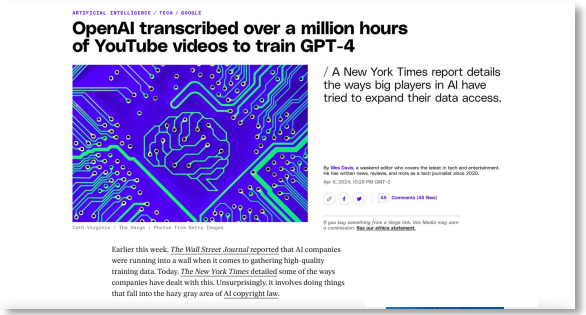
- Buy or access data products to develop better services (including LLMs)
  - Multilingual data
  - Multimodal data
  - Domain-specific data
  - All European languages
  - Easy discoverability and access
- Limited effort: keep existing infrastructure
- Legal compliance by design
  - Compliance with EU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


REUTERS
Exclusive: Reddit in AI content licensing deal with Google
By Anna Tong, Echo Wang and Martin Coulter
reddit


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.
As news publishers ink deals with AI companies to train their models with news 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
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
Gift this article
Save
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.
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.
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.
As 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.

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

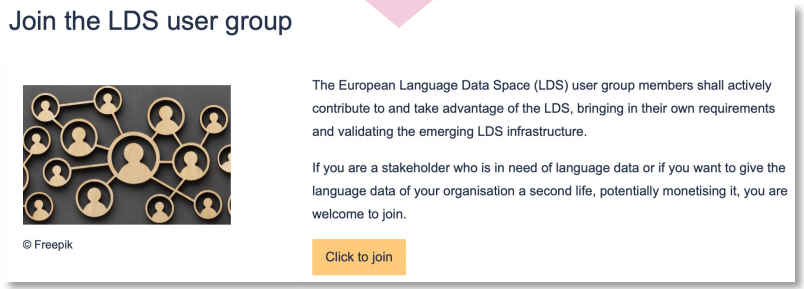
- A few examples of recent agreements:
  - Reddit: $60 million per year (Google)
  - Shutterstock: $25-50 million (Apple)
  - Springer: Tens of millions (Open AI)
  - Offer for news publishers: $1-5 million per year (Open AI)
  - Offer for owners of large datasets: $50 million (Apple)
- Global market 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
- 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

- 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

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



- **Meetings and Conferences**
  - Inaugural meeting in March 2024
  - Second meeting in June 2024
  - Third Meeting yesterday
  - Launch Conference: early 2025
- **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 Q4 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.

Common European Language Data Space

Thank you!

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

Dr. Simon Ostermann (DFKI GmbH, Germany) – LDS Project Manager and Site Manager Saarbrücken
simon.ostermann@dfki.de

07-11-2024 LDS Country Workshop Slovakia
https://language-data-space.ec.europa.eu