EUROPEAN LANGUAGE DATA SPACE

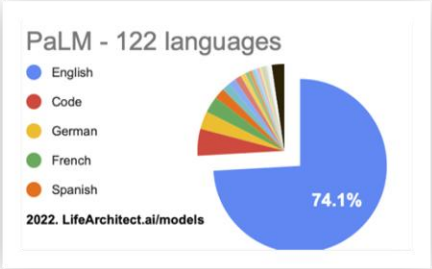
Country Workshop Malta

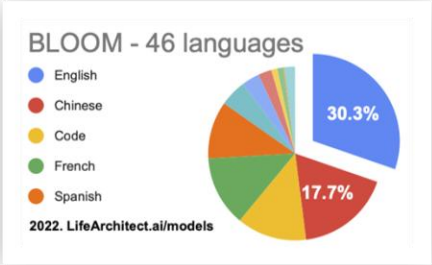
Khalid CHOUKRI (ELDA) choukri@elda.org

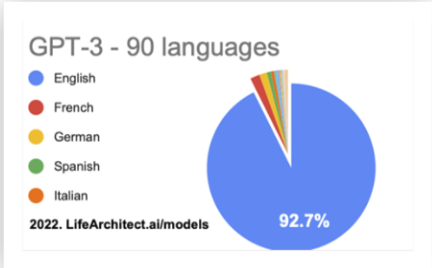
LDS Country Workshop in Malta
https://language-data-space.ec.europa.eu

Context: Large Language Models (LLMs)

- Unprecedented capabilities: Large language models are the most disruptive breakthrough in AI in recent history (GPT-3, ChatGPT, GPT-4, Claude, Gemini etc.)
- LLMs are trained on vast amounts of data and optionally also image, video, audio etc. data, i.e., multimodal data
- 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 …
- … and to make a difference with European data and European stakeholders.
- Already now billions and billions are made but …







LLM News

The Washington Post
Democracy Dies in Darkness

Subscribe Sign in

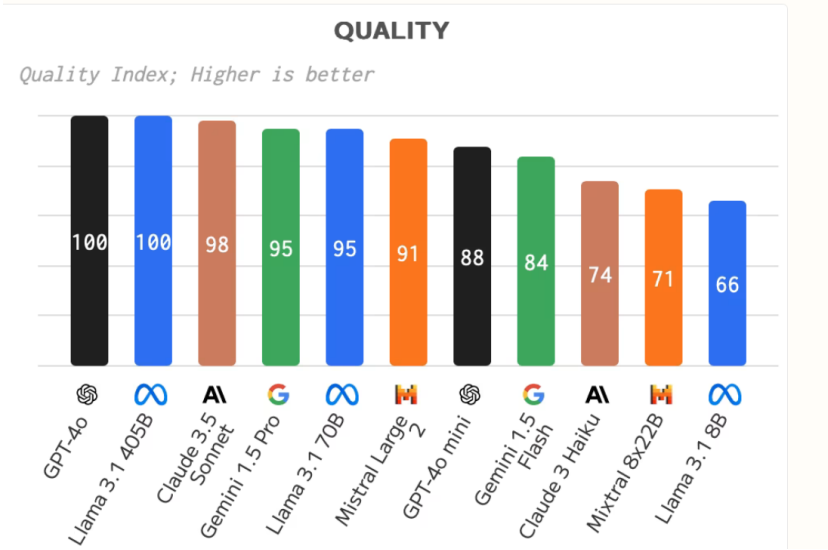
BUSINESS

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.

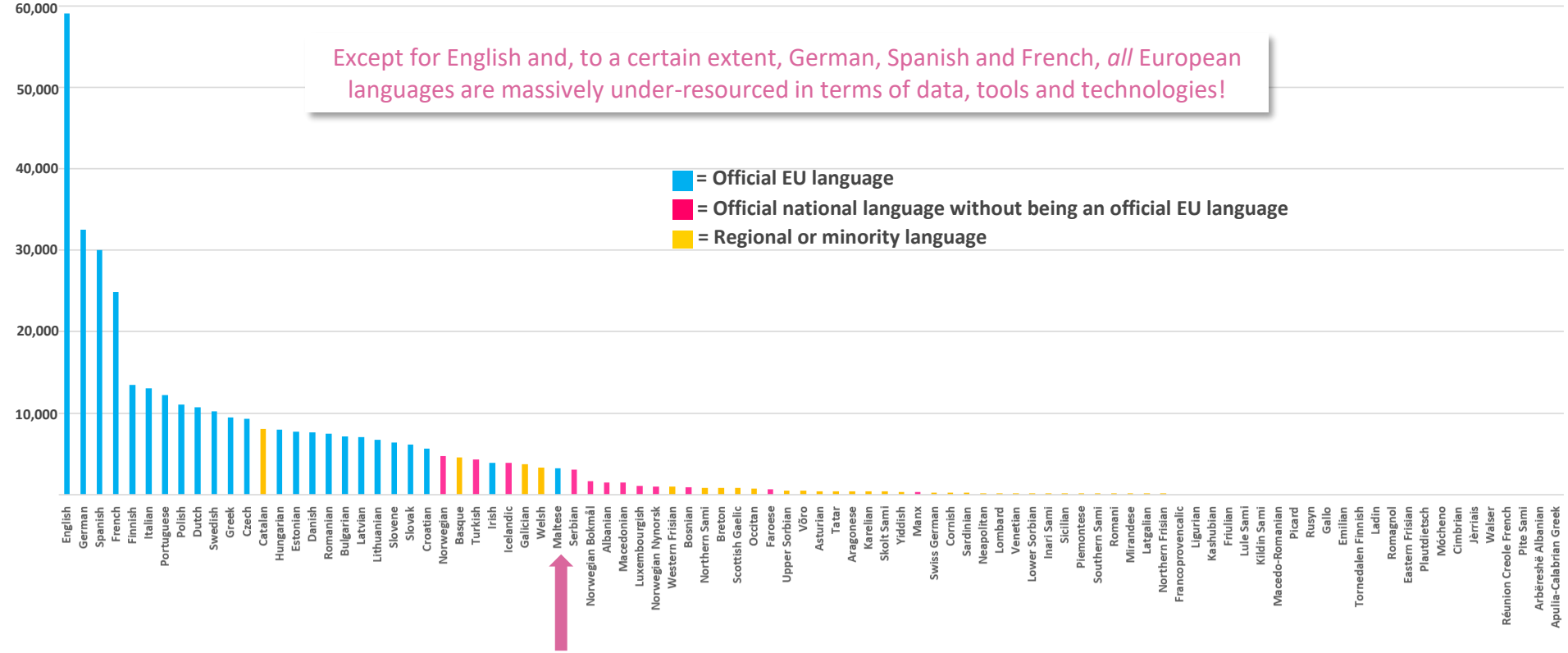


European Initiatives

- European initiatives for the development of LLMs
  - Large research projects in almost every country, e.g., Poland, Slovenia, Spain, Denmark, Italy, Germany etc.
  - Companies in many countries, e.g., Finland (Silo.ai), France (Mistral), Germany (Aleph Alpha)
  - EU-funded projects, e.g., HPLT, TrustLLM, OpenEuroLLM, LLMs4EU
  - New pan-European initiative: ALT-EDIC with an EC Support Action ALT-EDIC4EU
  - New EU initiative: AI Factories (tightly coupled with national HPC centres and EuroHPC JU)
- Challenges:
  - HPC facilities (amount, access and ease-of-use)
  - Speed of the big tech players in the US and Asia vs. speed of Europe
  - Availability of data for *all* European languages – *right now the most crucial bottleneck by far*

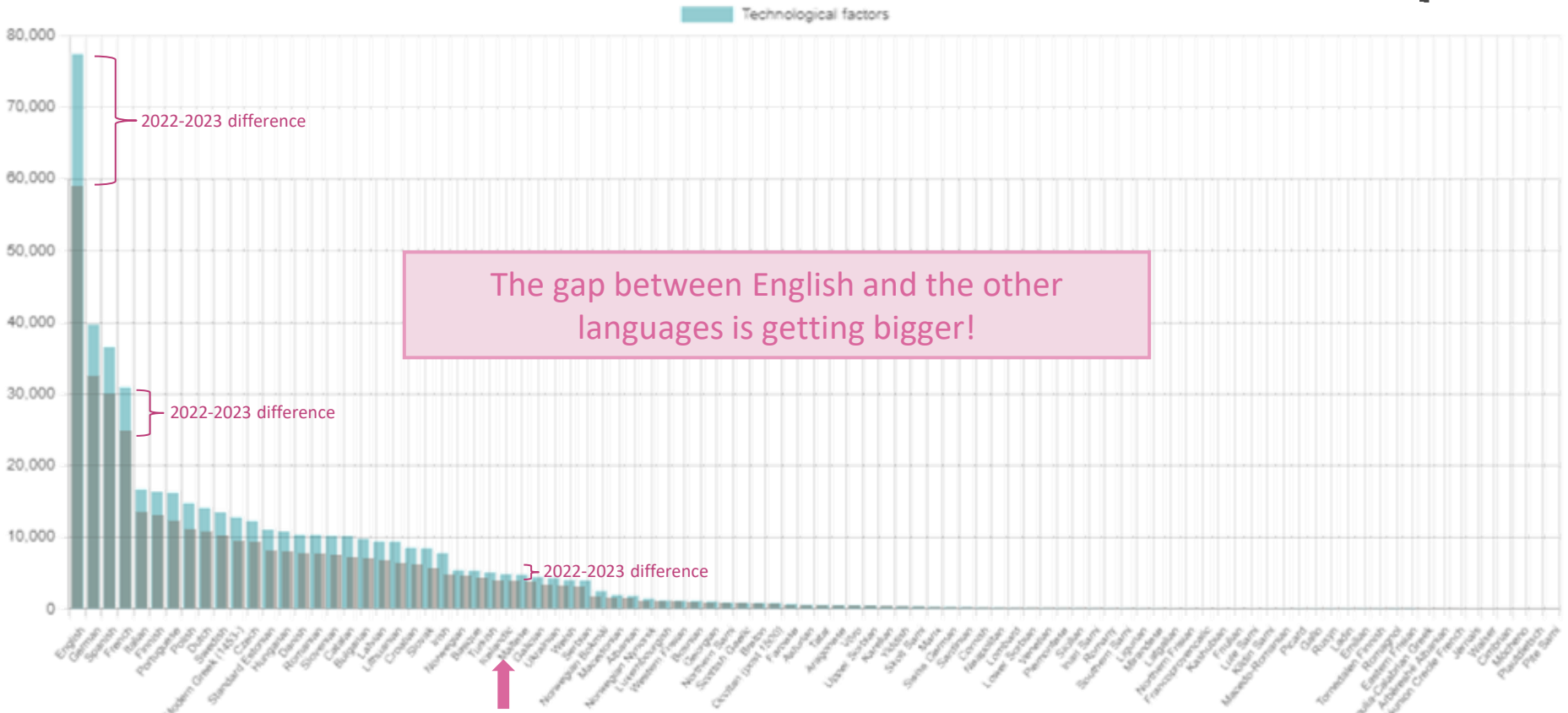
Digital Language Equality (DLE) Metric: Technological Scores

EUROPEAN LANGUAGE EQUALITY



DLE Metric: 2022 vs. 2023

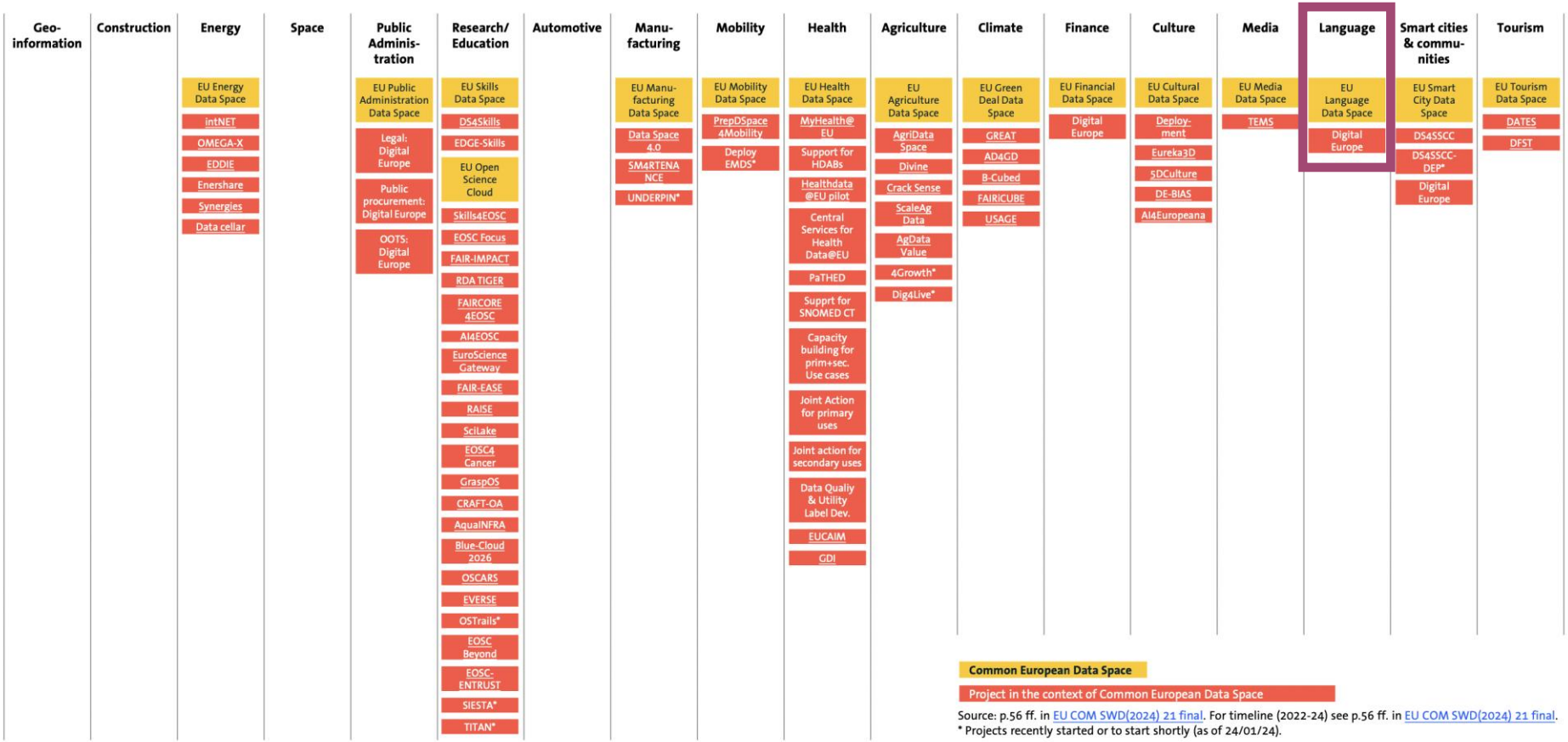
EUROPEAN LANGUAGE EQUALITY



EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy
- Data Spaces will help establish and grow the data economy in Europe
  - Organisations can sell and monetise their data, i.e., they can benefit from its value
- Convergence of the data economy initiatives in Europe (Gaia-X, IDSA, FIWARE, BDVA).
- Important initiatives in Europe:
  - DSSC – blueprint and specification development, community coordination and events
  - Simpl – development of Open Source data space components
  - Approx. 20++ data space initiatives with the EU's official mandate
    - The Common European Language Data Space is one of the official EU data space projects with a strong focus on industry
- European investment (both national and EU) already more than 2 billion Euros

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

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 trustworthy and secure European infrastructure and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- LDS is a marketplace for all organisations – commercial and public.
- LDS provides helpdesks for legal and for technical questions.
- 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.
- Technical development informed by ELG, ELRC-SHARE, META-SHARE.

Lead Partner and Coordinator
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Partners and Operation Leads
R.C. “Athena”, Institute for Language and Speech Processing
Evaluations and Language Resources Distribution Agency
TILDE
Main Subcontractors
3pc GmbH Neue Kommunikation
CLARIN ERIC
Big Data Value Association (Data, AI and Robotics) AISBL

LDS Launch Conference – 19 March 2025 – Villers-Cotterêts, France

Alliance for Language Technologies EDIC (ALT-EDIC)

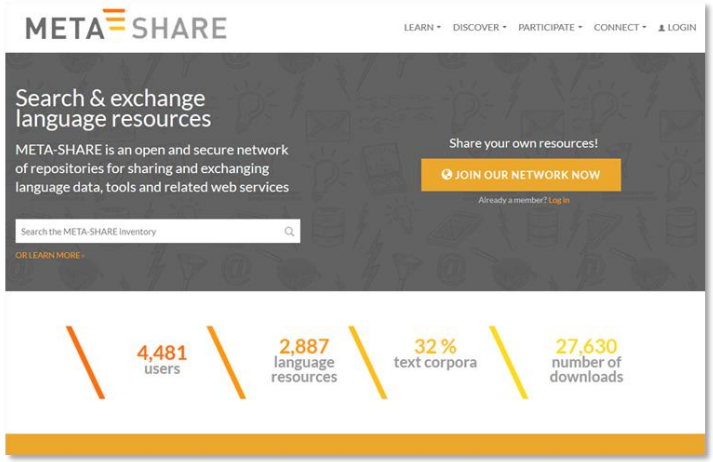
- European Digital Infrastructure Consortium (EDIC): a new legal entity type in the EU
- The first couple of EDICs are currently under development including the ALT-EDIC
- Coordinated by the French Ministry of Culture
- Close collaboration between: ALT-EDIC Working Group, EC, LDS
- ALT-EDIC action plan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

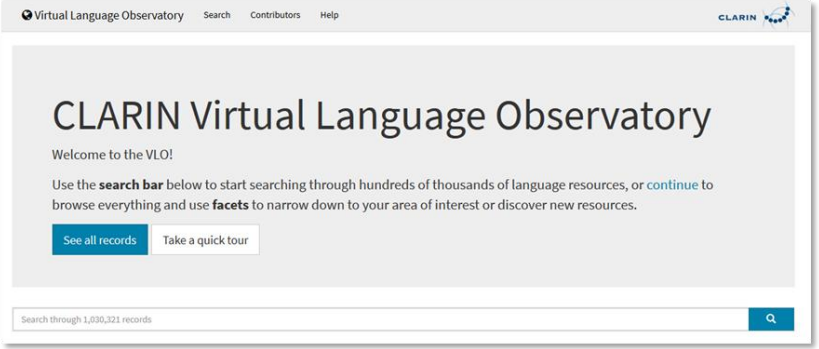
ALT-EDIC Members

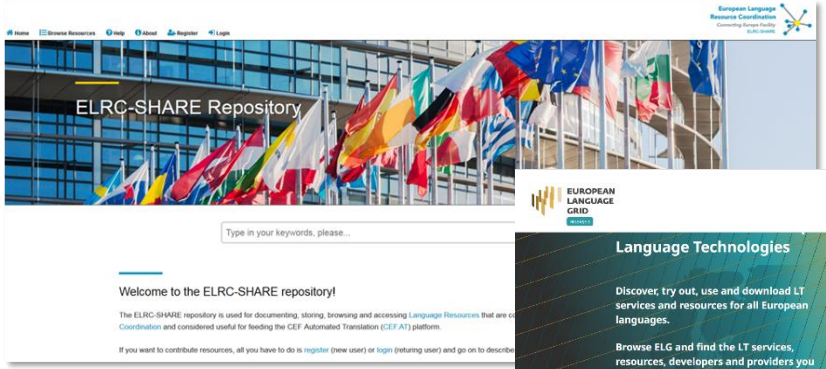
17 Members States: Bulgaria, Croatia, Czechia, Denmark, Finland, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Slovenia and Spain;

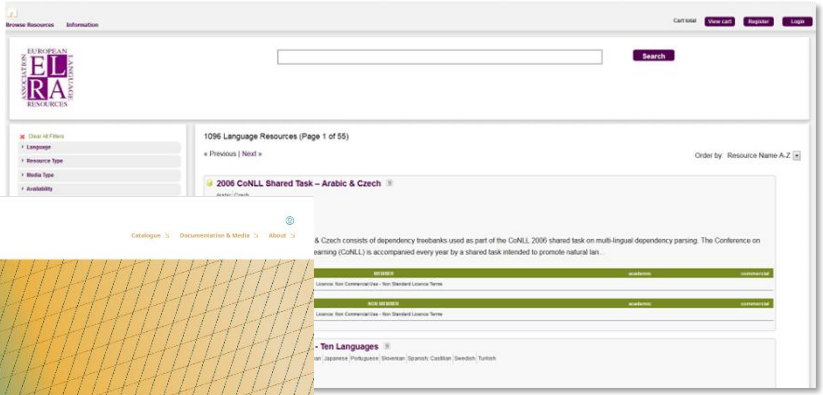
9 observing Member States: Austria, Belgium, Cyprus, Estonia, **Malta**, Portugal, Romania and Slovakia, and the Flanders (region).

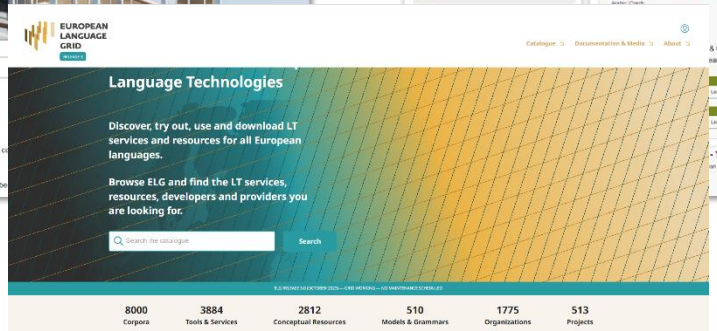
Long History of Language Data Sharing









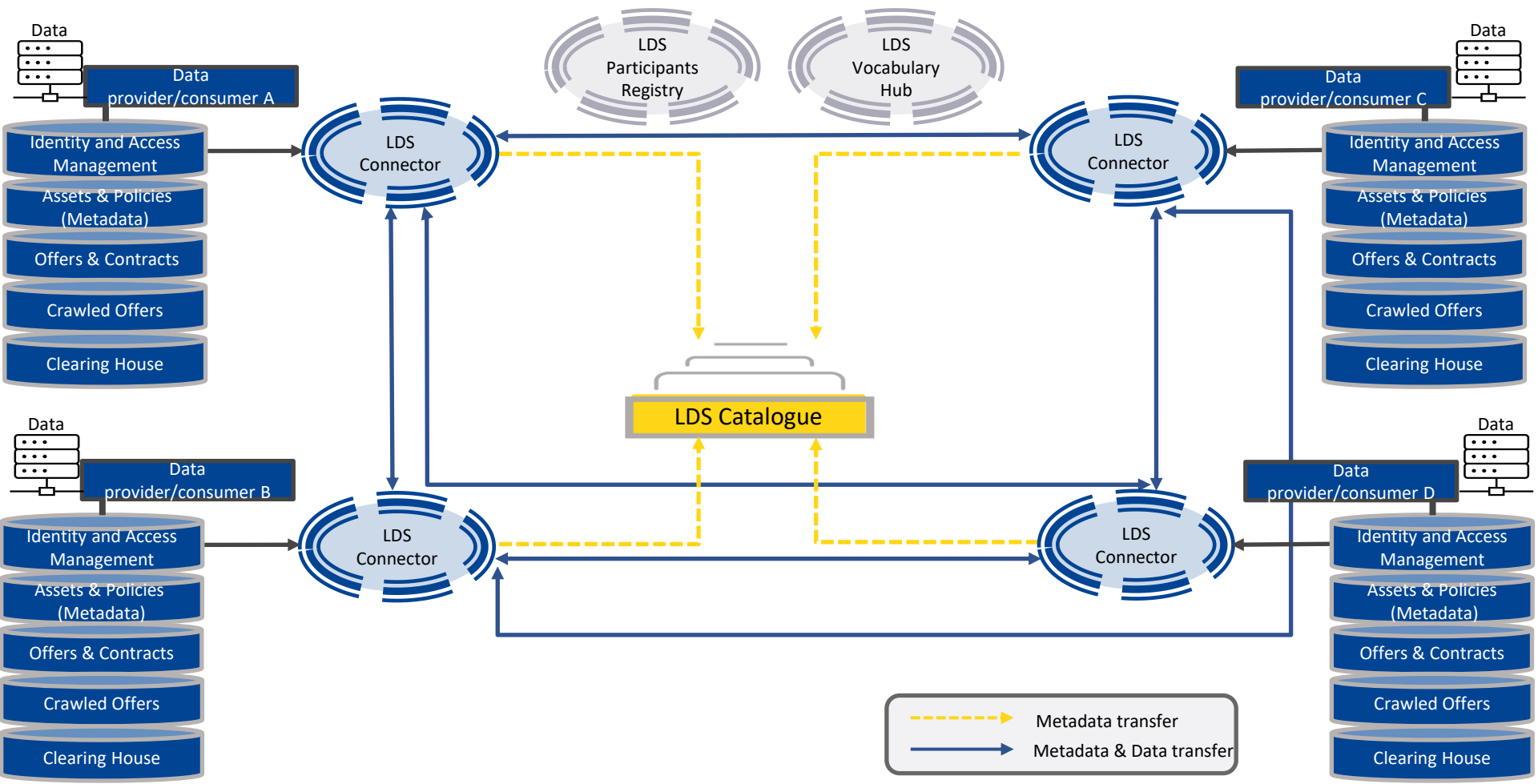


Classes of Data

Class of Data	Typical Size	Providers	Integration into LDS	Relevance, especially for LLM Development
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	High quality data; domain-specific data; data covering specific languages; raw data; processed data; evaluation data; data relevant for LLM development etc.

LDS Infrastructure – Basic Functionalities

- **Apply**, i.e., **become a participant** in the LDS marketplace
- **Discover** data offers (= data products) through the joint Language Data Space (LDS) catalogue
- **Negotiate access** to a data product
- **Transfer** data product from provider to consumer
- **Provide** data products


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

Typical Workflow – Purchase Process

- **Data provider** prepares a new data asset and adds a license (obligatory) and policy (optional)
- **Data provider** asks LDS helpdesk for advice regarding certain technical or legal questions
- **Data provider** publishes an offer through their LDS Connector
- This offer is now visible throughout the whole LDS marketplace, i.e., in all LDS Connector instances
- **Data consumers** can now discover this new offer through their own LDS Connector
- **Specific data consumer** evaluates the new offer …
- … and decides to *purchase the data offer* from the **data provider** through the LDS Connector
- The corresponding data set is transferred from the provider's connector to the buyer's connector

Language Data Space Marketplace – Value Proposition

Data Providers

- Sell data products (= data sets, data offers)
  - 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 AI-based 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

Language Data – Language Resources – Data Products – Data Offers

- 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: identify and make available, through the LDS, 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
    - Many unique LRs developed by European research organisations were licensed by European but also large US tech companies, e.g., for online NLP services (Machine Translation)

Futurism

TAP RUNNETH DRY | 11.13.23, 4:05 PM EST *by* MAGGIE HARRISON DUPRÉ

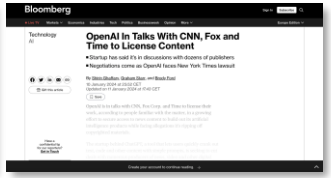
AI Companies Are Running Out of Training Data

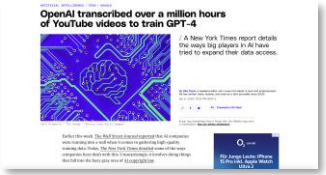
The well is running dry.

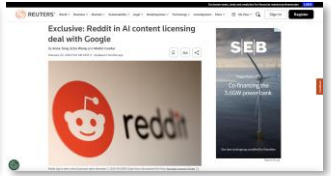
/ Artificial Intelligence / Ai / Ai Industry / Ai Training

Data Products expected in LDS – Training Data for Gen. AI and LLMs

- A few examples of recent data 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
- It's up to the data providers to establish offers and prices that make sense for them
- Our ambition is to establish LDS as *the* marketplace for European language data







Technical Helpdesk – Legal Helpdesk – Business Development Helpdesk

- **Legal helpdesk:** provides legal advice and assistance to all stakeholders involved in the LDS. It provides guidance for legal questions related to the collection and sharing of language data, licensing schemes, IPR clearance, data protection requirements as well as confidentiality aspects. These may include:
  - questions related to any kind of utilisation of language data in the LDS;
  - legal queries, e.g., how to exploit the LDS by clearing specific legal aspects, etc.

- **Technical helpdesk:** provides technical advice and assistance to all stakeholders involved in the LDS, providing first-line support to technical questions. These may include:
  - general troubleshooting and assistance on the usage of the LDS;
  - technical support, e.g., validation of technical compliance of data assets

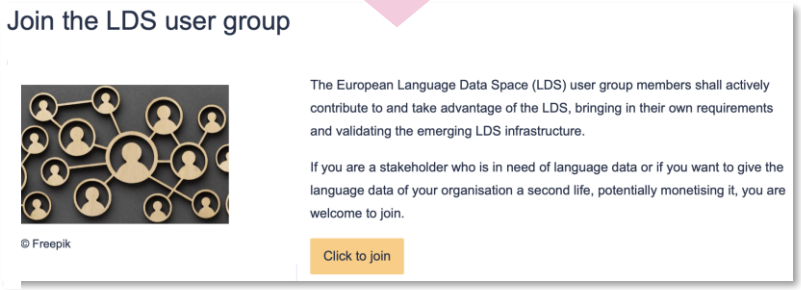
- **Business Development helpdesk** helps monetise language data with experts in business development, pricing and revenues strategies, licensing issues for valuable assets, etc.

Next Steps

- LDS is in full swing: technical development, promotion, dissemination, governance etc.
- Testing: LDS software, metadata model, license and policy templates
- Adoption of LDS by industry and other organisations → grow the LDS User Group
- Collaborations with
  - DSSC, Simpl and ALT-EDIC
  - European projects, e.g., HPLT, OpenGPT-X, OpenWebSearch
  - Other relevant data spaces, especially Media (TEMS) and Cultural Heritage (Europeana)
  - EuroHPC Joint Undertaking
- Identify and make available new, fresh, interesting, novel, relevant 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 in November 2024
  - The LDS User Group is constantly growing
  - Testing the LDS connector (currently 10+ instances)

LDS User Group (as of 05 March 2025)	
Individual Members	Unique Organisations
198	158

- **Please join the LDS User Group or encourage members of your networks to join!**
  - Validation of concepts, ideas, software; first test installations of the LDS connector (*currently ongoing!*); first trial exchanges of data; surveys etc.

Common European Language Data Space

Thank you!

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

Khalid CHOUKRI (ELDA) choukri@elda.org

LDS Country Workshop in Malta
https://language-data-space.ec.europa.eu