EUROPEAN LANGUAGE DATA SPACE

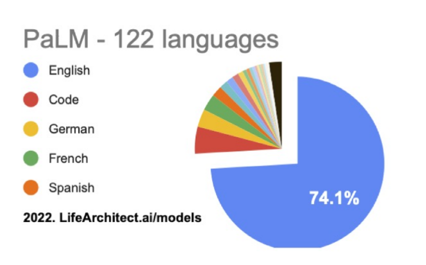
Country Workshop Poland

Prof. Dr. Georg Rehm, Katrin Marheinecke (DFKI GmbH, Germany) – LDS Coordination
(georg.rehm@dfki.de, katrin.marheinecke@dfki.de)

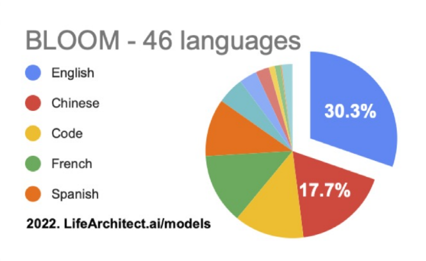
29-05-2025 LDS Country Workshop Poland
https://language-data-space.ec.europa.eu

Context: Large Language Models (LLMs)

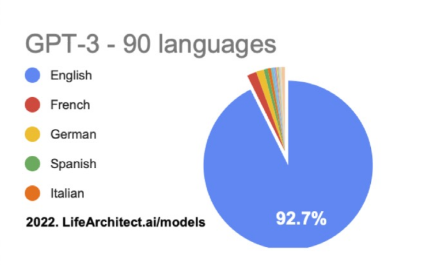
- Unprecedented capabilities: Large language models are the most disruptive breakthrough in AI in recent history (GPT-3, ChatGPT, GPT-4, Claude, Gemini etc.)
- LLMs are trained on vast amounts of data and optionally also image, video, audio etc. data, i.e., multimodal data
- Multilingualism makes everything 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 …



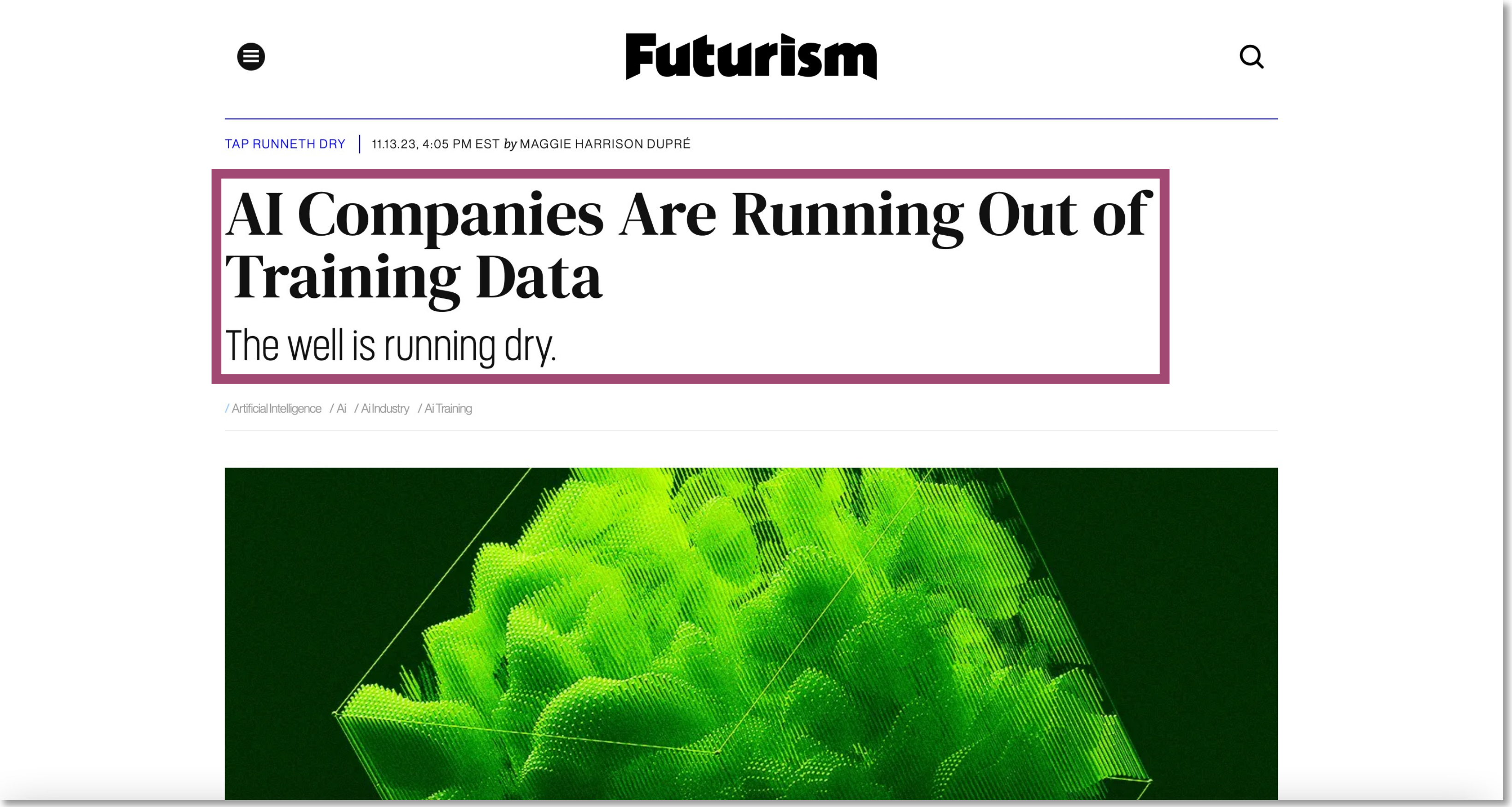
2022. LifeArchitect.ai/models



2022. LifeArchitect.ai/models



2022. LifeArchitect.ai/models



Futurism

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

AI Companies Are Running Out of Training Data

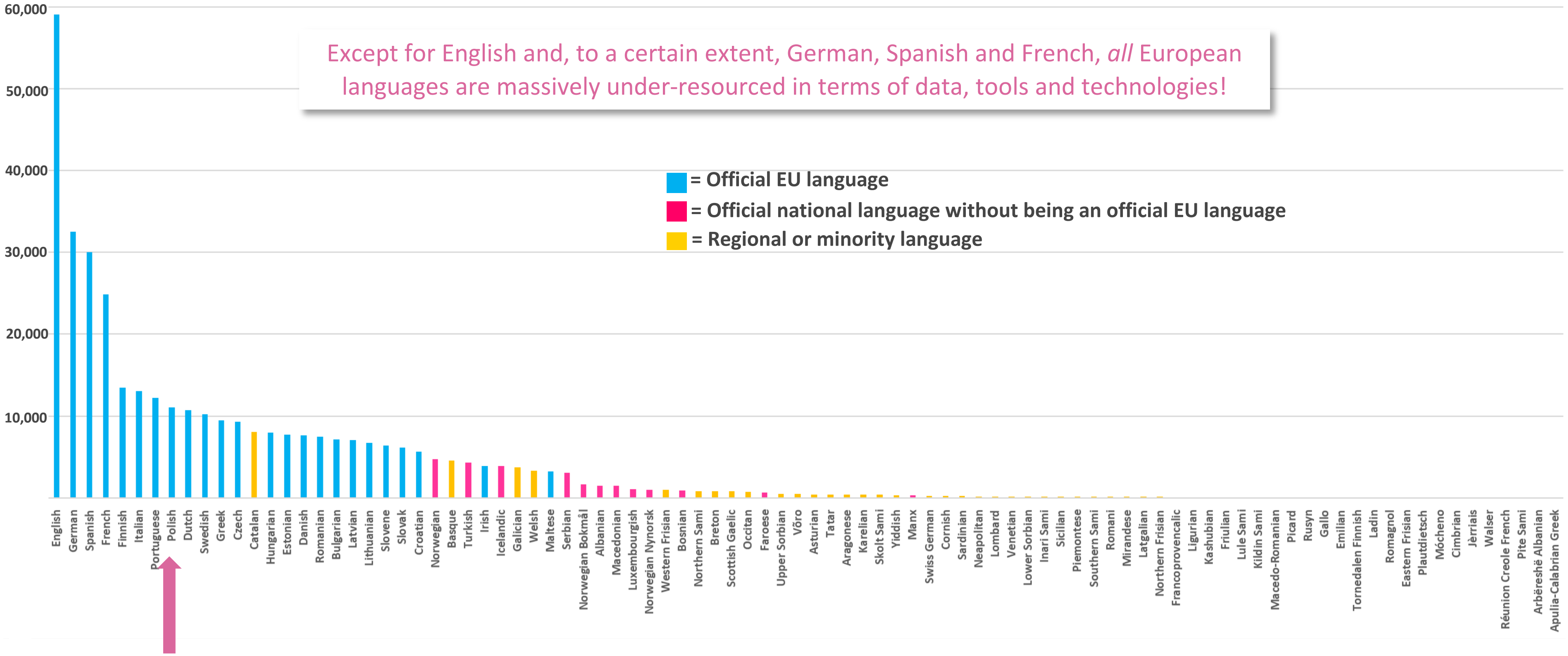
The well is running dry.

/ Artificial Intelligence / Ai / Ai Industry / Ai Training

European Initiatives

- European initiatives for the development of LLMs
  - Large research projects in almost every country, e.g., Spain, Poland, 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
  - 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 Metric: Technological Scores



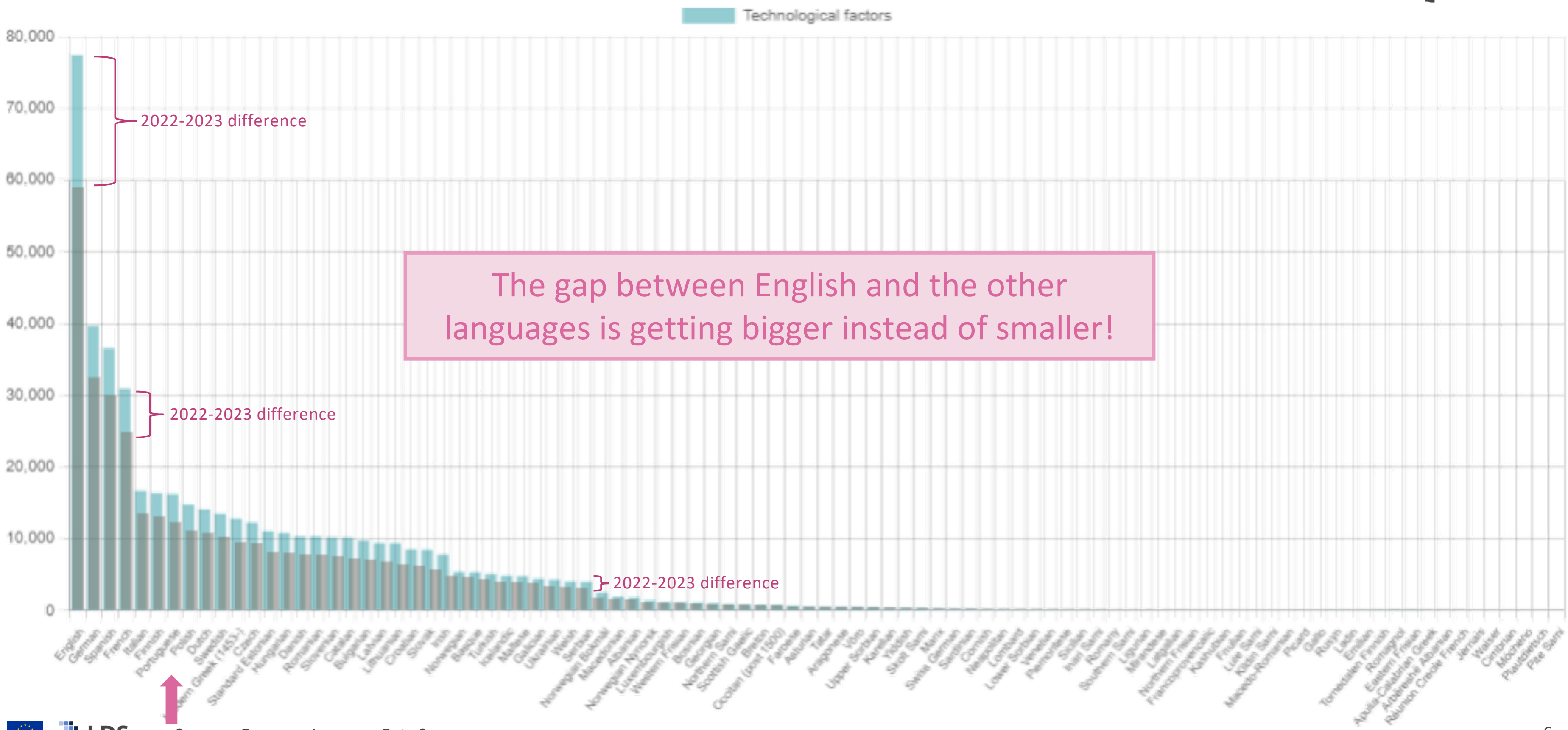
Except for English and, to a certain extent, German, Spanish and French, *all* European languages are massively under-resourced in terms of data, tools and technologies!

🟦 = Official EU language

🟥 = Official national language without being an official EU language

🟨 = Regional or minority language

DLE Metric: 2022 vs. 2023



2022-2023 difference

2022-2023 difference

2022-2023 difference

The gap between English and the other languages is getting bigger instead of smaller!

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

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.

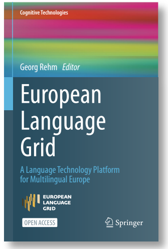
Lead Partner and Coordinator
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Partners and Operation Leads
R.C. “Athena”, Institute for Language and Speech Processing
Evaluations and Language Resources Distribution Agency
TILDE
Main Subcontractors
3pc GmbH Neue Kommunikation
CLARIN ERIC
Big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

META-NET

European Language Resource Coordination
Connecting Europe Facility

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY



Cognitive Technologies

Georg Rehm *Editor*

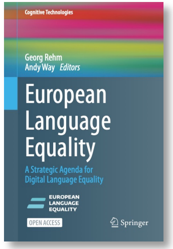
European Language Grid

A Language Technology Platform for Multilingual Europe

EUROPEAN LANGUAGE GRID

OPEN ACCESS

Springer



Cognitive Technologies

Georg Rehm
Andy Way *Editors*

European Language Equality

A Strategic Agenda for Digital Language Equality

EUROPEAN LANGUAGE EQUALITY

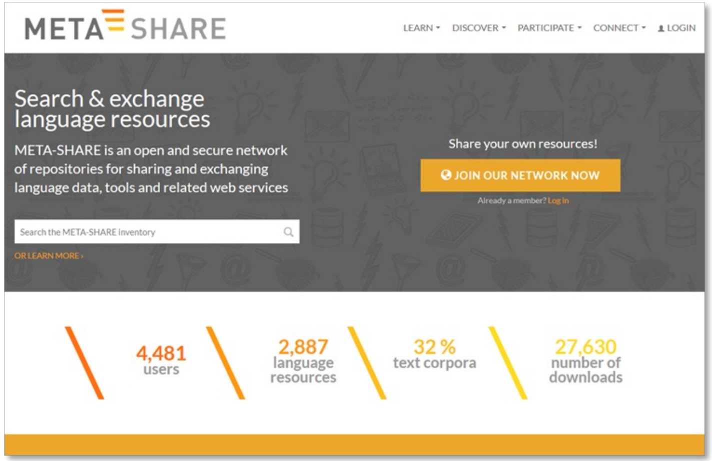
OPEN ACCESS

Springer

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

EUROPEAN
LANGUAGE
DATA SPACE
Launch Conference
19-03-2025 LDS Launch Conference, Villers-Cotterêts, France
https://language-data-space.ec.europa.eu
Rechercher
European Commission
EUROPEAN
LANGUAGE
DATA SPACE
Share. Connect. Benefit.
The European marketplace for language data.
https://language-data-space.ec.europa.eu
The European Language Data Space is a service contract (no. LC_01936389) operating under the DIGITAL Work Programme 2021-2022.

Long History of Language Data Sharing



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

Search the META-SHARE inventory

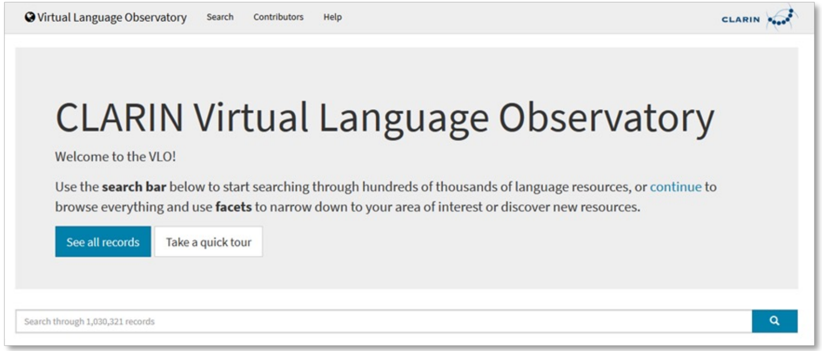
OR LEARN MORE ›

Share your own resources!

JOIN OUR NETWORK NOW

Already a member? Log in

4,481	2,887	32 %	27,630
users	language resources	text corpora	number of downloads



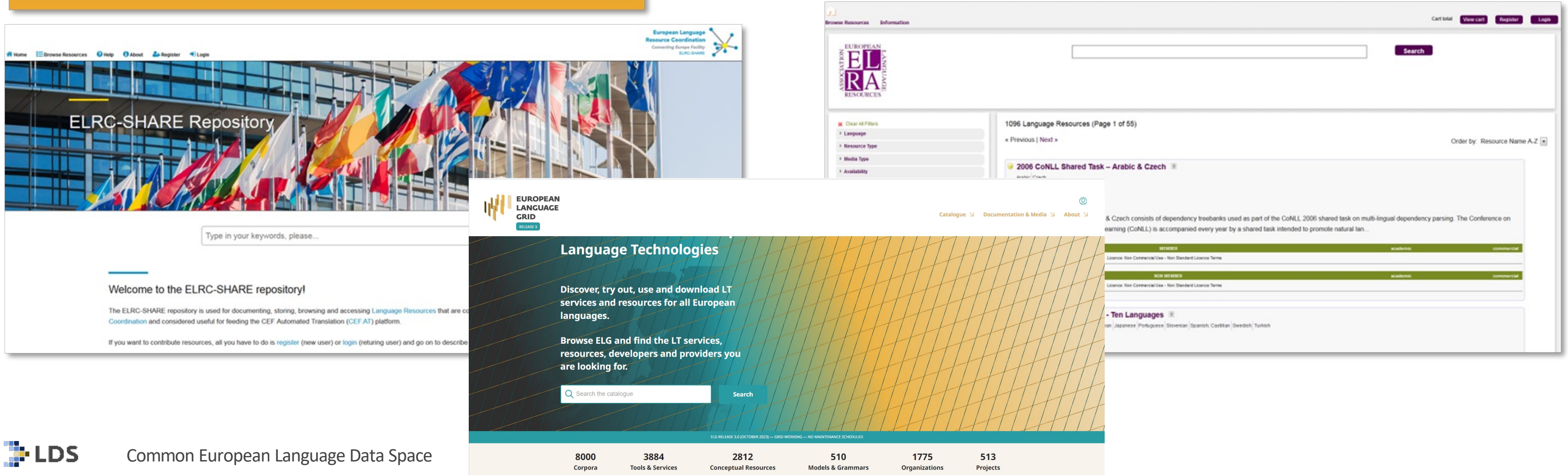
CLARIN Virtual Language Observatory

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



Home
Browse Resources
Help
About
Register
Login
European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

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 c
Coordination 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 (returing user) and go on to describe

1096 Language Resources (Page 1 of 55)

« Previous | Next »

Order by: Resource Name A-Z

2006 CoNLL Shared Task – Arabic & Czech

Arabic | Czech

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

MEMBER	academic	commercial
Licence: Non Commercial Use - Non Standard Licence Terms		

NON MEMBER	academic	commercial
Licence: Non Commercial Use - Non Standard Licence Terms		

- Ten Languages

an | Japanese | Portuguese | Slovenian | Spanish; Castilian | Swedish | Turkish

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 | Search

ELG RELEASE 3.0 (OCTOBER 2023) — GRID WORKING — NO MAINTENANCE SCHEDULED

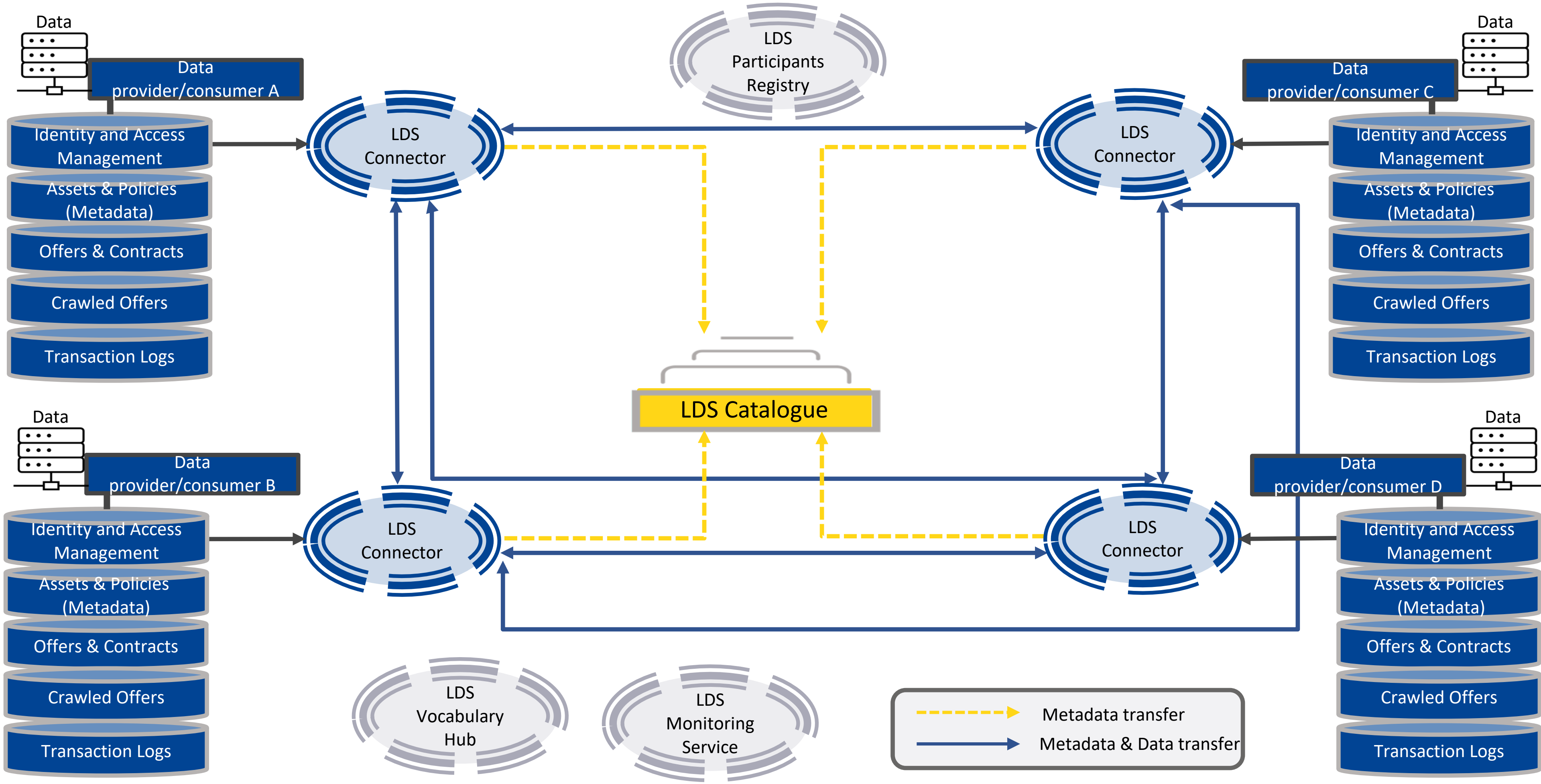
8000	3884	2812	510	1775	513
Corpora	Tools & Services	Conceptual Resources	Models & Grammars	Organizations	Projects

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)

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.



Data

Data provider/consumer A

Identity and Access Management

Assets & Policies (Metadata)

Offers & Contracts

Crawled Offers

Transaction Logs

LDS Connector

LDS Participants Registry

Data

Data provider/consumer C

Identity and Access Management

Assets & Policies (Metadata)

Offers & Contracts

Crawled Offers

Transaction Logs

LDS Connector

LDS Catalogue

Data

Data provider/consumer B

Identity and Access Management

Assets & Policies (Metadata)

Offers & Contracts

Crawled Offers

Transaction Logs

LDS Connector

Data

Data provider/consumer D

Identity and Access Management

Assets & Policies (Metadata)

Offers & Contracts

Crawled Offers

Transaction Logs

LDS Connector

LDS Vocabulary Hub

LDS Monitoring Service

Metadata transfer

Metadata & Data transfer

LDS Infrastructure – Basic Functionalities

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

Typical Workflow – Purchase Process: Providers

- **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
  - Find the process summarised in this YouTube video

Typical Workflow – Purchase Process: Providers

LEARN MORE

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

Typical Workflow – Purchase Process: Consumers

- **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
  - Find the process summarised in this YouTube video

You will find more practical guidance on http://www.youtube.com/@LangDataSpace

and in the LDS User Guide at https://docs.language-data-space.eu

Typical Workflow – Purchase Process: Consumers

LEARN MORE

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

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

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** (under development): helps monetise language data

Data Quality

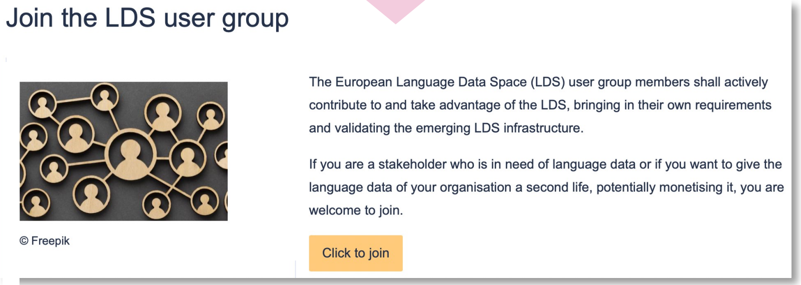
- FAQ: “What about data quality? Do you measure or enforce a certain quality level?”
  - No, we don’t.
  - Reason: the quality of language data cannot be measured in a generic way.
- Nota bene: this is different in other domains, e.g., sensor data, where there are various quality-related aspects, e.g., resolution, continuity etc.
- For one person and use case, a specific data set can be very high quality and extremely useful and for another person, the same data set can be useless.
- LDS relies on meaningful metadata records that provide sufficient detail about a certain dataset.
- Goal: the quality or substance of a dataset should be conveyed through its metadata record.

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
  - Data Spaces Support Centre (DSSC), Simpl and ALT-EDIC
  - European projects, e.g., HPLT, Hivemind, OpenWebSearch
  - Other relevant data spaces, especially Media, Tourism and Cultural Heritage (Europeana)
  - EuroHPC Joint Undertaking (PPP)
- 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



Join the LDS user group

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.

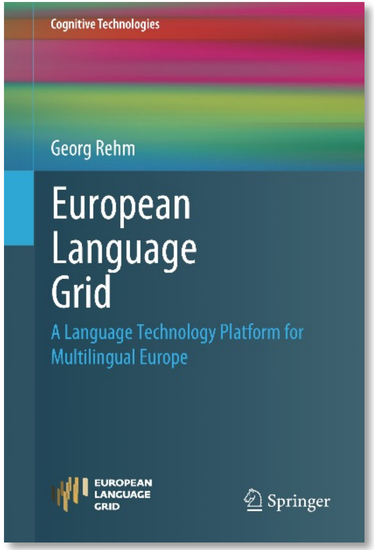
- **Meetings and Conferences**
  - Inaugural meeting in March 2024
  - Second meeting in June 2024
  - Third meeting in November 2024
  - Next meeting planned for fall 2025 (next release)
  - The LDS User Group is constantly growing
  - Testing the LDS connector (currently 15 instances)

LDS User Group (as of 27 May 2025)	
Individual Members	Unique Organisations
244	190

- **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!


Cognitive Technologies

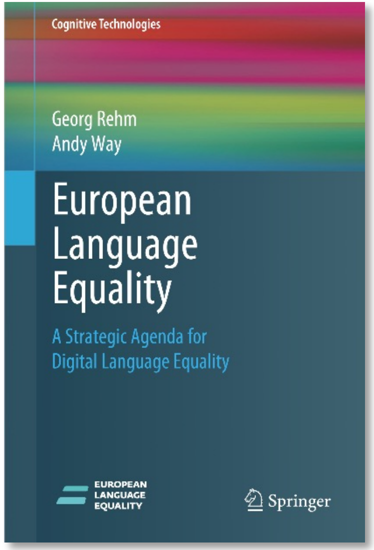
Georg Rehm

European Language Grid

A Language Technology Platform for Multilingual Europe

EUROPEAN LANGUAGE GRID

Springer


Cognitive Technologies

Georg Rehm
Andy Way

European Language Equality

A Strategic Agenda for Digital Language Equality

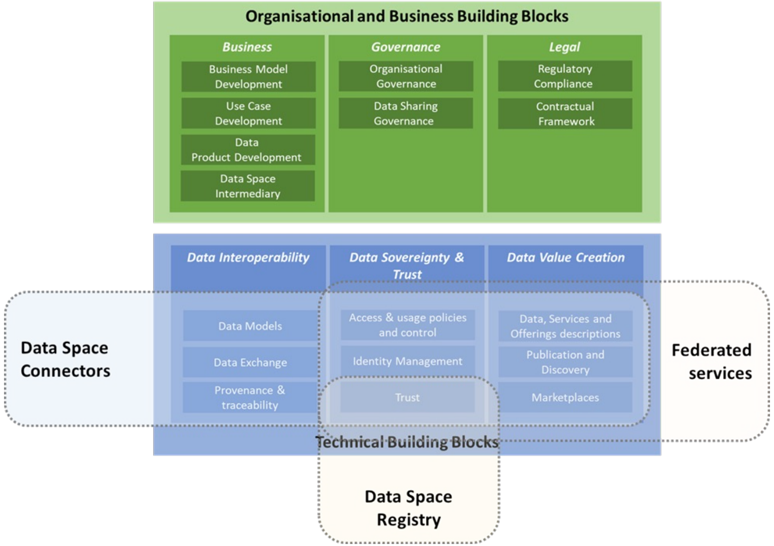
EUROPEAN LANGUAGE EQUALITY

Springer

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

Prof. Dr. Georg Rehm, Katrin Marheinecke (DFKI GmbH, Germany) – LDS Coordination
coordination@language-data-space.eu

29-05-2025 LDS Country Workshop Poland
https://language-data-space.ec.europa.eu



Organisational and Business Building Blocks

Business
- Business Model Development
- Use Case Development
- Data Product Development
- Data Space Intermediary

Governance
- Organisational Governance
- Data Sharing Governance

Legal
- Regulatory Compliance
- Contractual Framework

Technical Building Blocks

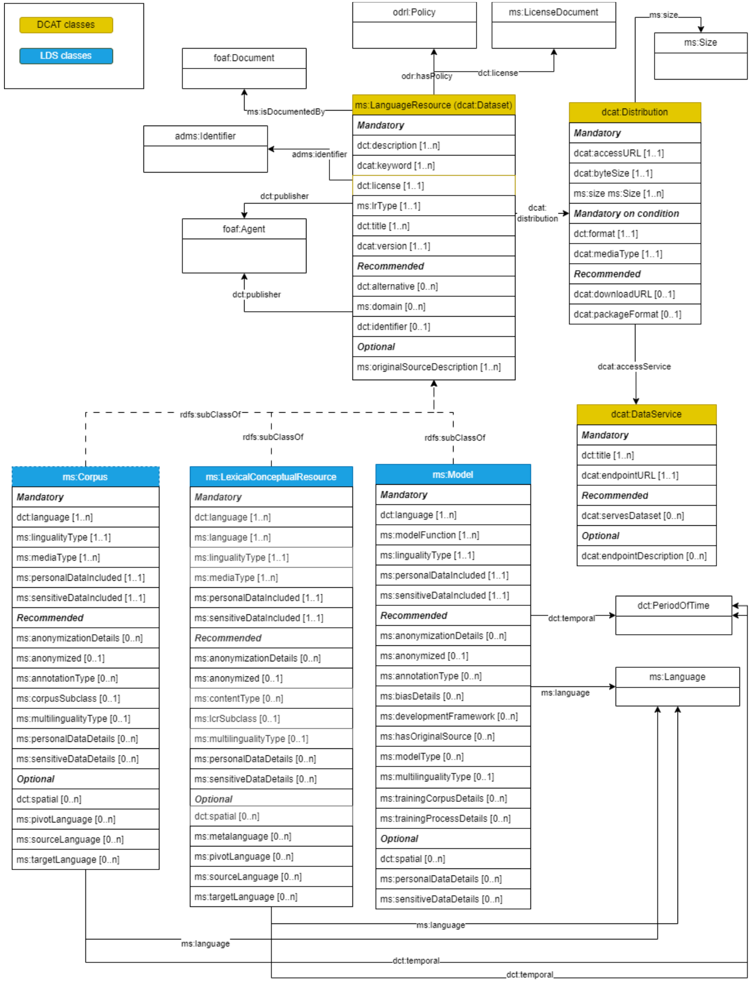
Data Interoperability
- Data Models
- Data Exchange
- Provenance & traceability

Data Sovereignty & Trust
- Access & usage policies and control
- Identity Management
- Trust

Data Value Creation
- Data, Services and Offerings descriptions
- Publication and Discovery
- Marketplaces

Built on Existing Solutions

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



DCAT classes

LDS classes

ms:LanguageResource (dcat:Dataset)
Mandatory
dct:description [1..n]
dcat:keyword [1..n]
dct:license [1..1]
ms:lrType [1..1]
dct:title [1..n]
dcat:version [1..1]
Recommended
dct:alternative [0..n]
ms:domain [0..n]
dct:identifier [0..1]
Optional
ms:originalSourceDescription [1..n]

dcat:Distribution
Mandatory
dcat:accessURL [1..1]
dcat:byteSize [1..1]
ms:size ms:Size [1..n]
Mandatory on condition
dct:format [1..1]
dcat:mediaType [1..1]
Recommended
dcat:downloadURL [0..1]
dcat:packageFormat [0..1]

dcat:DataService
Mandatory
dct:title [1..n]
dcat:endpointURL [1..1]
Recommended
dcat:servesDataset [0..n]
Optional
dcat:endpointDescription [0..n]

ms:Corpus
Mandatory
dct:language [1..n]
ms:lingualityType [1..1]
ms:mediaType [1..n]
ms:personalDataIncluded [1..1]
ms:sensitiveDataIncluded [1..1]
Recommended
ms:anonymizationDetails [0..n]
ms:anonymized [0..1]
ms:annotationType [0..n]
ms:corpusSubclass [0..1]
ms:multilingualityType [0..1]
ms:personalDataDetails [0..n]
ms:sensitiveDataDetails [0..n]
Optional
dct:spatial [0..n]
ms:pivotLanguage [0..n]
ms:sourceLanguage [0..n]
ms:targetLanguage [0..n]

ms:LexicalConceptualResource
Mandatory
dct:language [1..n]
ms:language [1..n]
ms:lingualityType [1..1]
ms:mediaType [1..n]
ms:personalDataIncluded [1..1]
ms:sensitiveDataIncluded [1..1]
Recommended
ms:anonymizationDetails [0..n]
ms:anonymized [0..1]
ms:contentType [0..n]
ms:lcrSubclass [0..1]
ms:multilingualityType [0..1]
ms:personalDataDetails [0..n]
ms:sensitiveDataDetails [0..n]
Optional
dct:spatial [0..n]
ms:metalanguage [0..n]
ms:pivotLanguage [0..n]
ms:sourceLanguage [0..n]
ms:targetLanguage [0..n]

ms:Model
Mandatory
dct:language [1..n]
ms:modelFunction [1..n]
ms:lingualityType [1..1]
ms:personalDataIncluded [1..1]
ms:sensitiveDataIncluded [1..1]
Recommended
ms:anonymizationDetails [0..n]
ms:anonymized [0..1]
ms:annotationType [0..n]
ms:biasDetails [0..n]
ms:developmentFramework [0..n]
ms:hasOriginalSource [0..n]
ms:modelType [0..n]
ms:multilingualityType [0..1]
ms:trainingCorpusDetails [0..n]
ms:trainingProcessDetails [0..n]
Optional
dct:spatial [0..n]
ms:personalDataDetails [0..n]
ms:sensitiveDataDetails [0..n]