The Digital Europe Programme and the Common European Language Data Space

LDS workshop – Belgium and the Netherlands

Language Technology for a Multilingual Europe

Mateusz Szoturma
Legal and Policy Officer at the European Commission (DG CONNECT – G3)

(Multilingual and European)

AI Revolution!

European Commission

Background

Current Landscape

- Advancement from non-EU
- Language Bias
- Trust Issues
- Cannot scale-up without data

EU Response

- Need for transparent, trustworthy ecosystem that supports Data control, language diversity and EU values

Background

Current Landscape

- Advancement from non-EU
- Language Bias
- Trust Issues
- Cannot scale-up without data

EU Response

- Need for transparent, trustworthy ecosystem that supports Data control, language diversity and EU values

EUROPEAN LANGUAGE DATA SPACE

ALT-EDIC

DIGITAL calls 2024

European Commission

The Common European Language Data Space (LDS)

- Overview
- Architecture

European Commission

Language Data Space – Overview

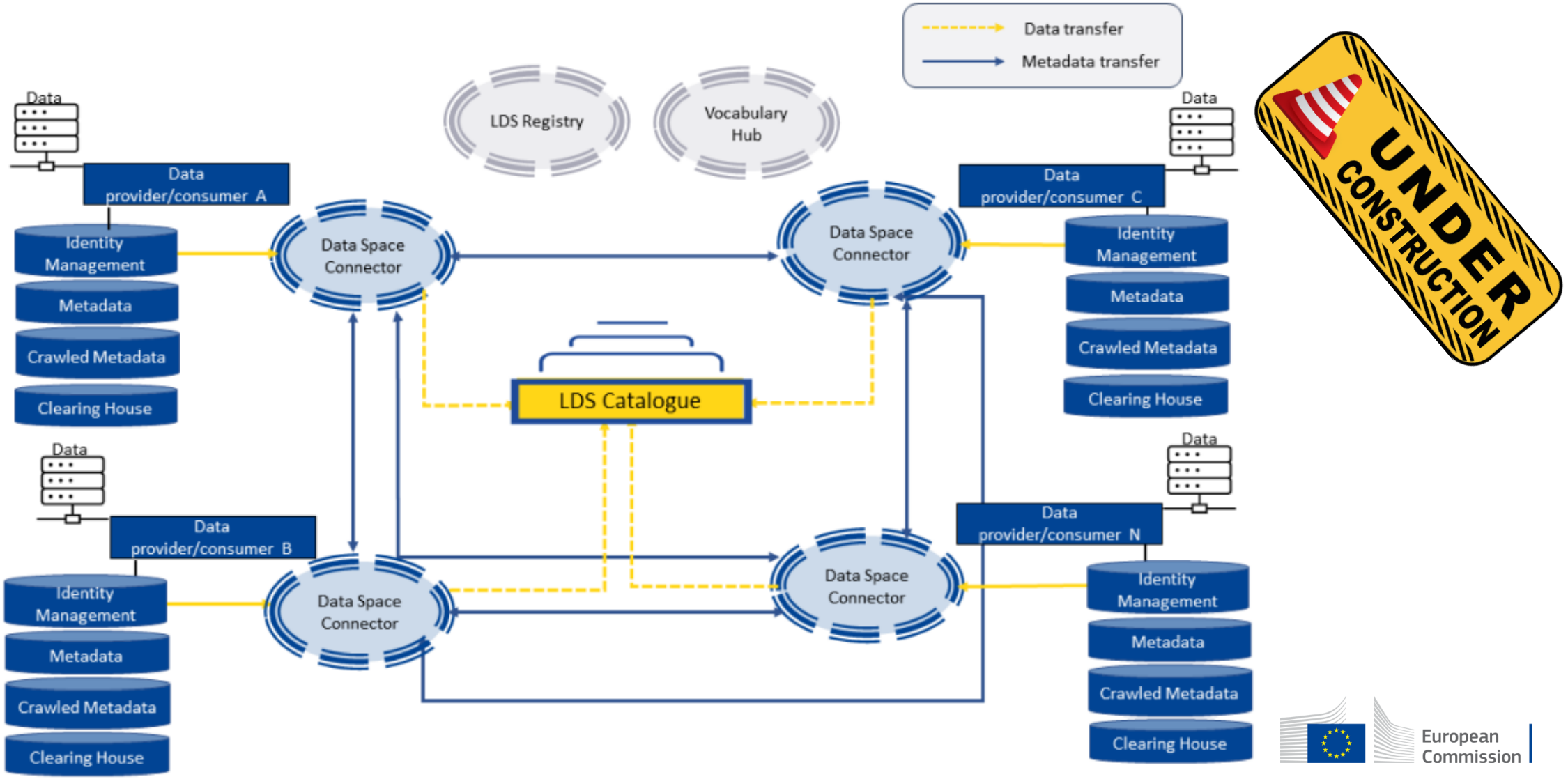
EUROPEAN LANGUAGE DATA SPACE

What	**Platform** and **marketplace** for the collection, creation, sharing and re-use of multilingual and multimodal language data	
	6 Mio € **Procurement** Contract from the **DIGITAL** Work Programme 2021-2022	
Why	Increase **availability** of clean, high-quality, compliant language data	
	Support the development of state-of-the-art European language technologies	
Who	**Consortium**: DFKI, ELDA, TILDE, Athena Institute	
	Governance: Member States representatives and Industry User group	
	Stakeholders: Industry, Public Administration, Research, etc.	
How	- Technical Architecture and Infrastructure	- Openess
	- Governance Framework	- Respect of EU Rules and Values
	- Promotion	- Overcoming Legal Barriers (IPR, GDPR, etc.)
When	Started in **January 2023** for Three years – until end 2025 (Possible contract renewal for another year).	
	First version expected in **Q4 2024**	

European Commission

Language Data Space – Architecture

EUROPEAN LANGUAGE DATA SPACE



European Commission

Alliance for Language Technology EDIC (ALT-EDIC)

European Commission

ALT-EDIC

Alliance for Language Technologies
European Digital Infrastructure Consortium

Objectives

- Technological leadership and strategic autonomy
- Preserve linguistic and cultural diversity in Europe
- Respect European rules and values
- Cooperation
- Raising awareness

Participation

Members [17]

FR, BG, CZ, DK, ES, FI, GR, HR, HU, IE, IT, LT, LV, LU, NL, PL, SI

Observers [8]

AT, BE, CY, EE, MT, PT, RO, SK

Interested [2]

DE, SE

BUDGET

50M€

Action Plan - Data

Data

Data Ecosystem

Support the development and training of Language Models (LMs)

Language Coverage

Ensures a **diverse** and **inclusive** linguistic landscape.
- All European Languages
- Addresses the scarcity of training data for specific languages (less spoken languages)

Federation of Resources

- Federate existing EU and Member States language and multimodal resources.
- Foster exchange and collection of high-quality language data

Synergies

- Leverage the capabilities of the Language Data Space.
- Synergies with Member States initiatives

Action Plan – Existing Models

Existing Models

Open-Source Repository

- Creation of a comprehensive repository of existing open-source language models.
- Facilitating reuse by industrial actors within the ALT-EDIC ecosystem.

Fine-Tuning Methods for SMEs

- Development of tailored methods for fine-tuning language models.
- Focus on ensuring accessibility for SMEs in AI development.

Methodologies for Evaluation

- Introduction of robust methodologies for evaluating and certifying language models.
- Emphasis on ensuring high standards and performance in language technologies.

Addressing Bias in NLP Models

Implementation of measures for fair evaluation, certification, and normalization

Action Plan – New Models

New Models

Seed fund for new Language Models

- ALT-EDIC operates as a central pool seed fund.
- Facilitates the aggregation of public and private resources for language model projects.

Public-Private Collaboration

- Encourages collaboration between public and private entities.
- Aims to leverage diverse resources for robust and innovative language model development.

Access to High-Performance Computing (HPC)

- Provides essential access to European High-Performance Computing (HPC) resources.

Project Launch and Development

Enables the launch and ongoing development of new Large Language Model projects

Action Plan – Ecosystem

Ecosystem

Advisory Hub

- ALT-EDIC functions as a central advisory point.
- Offers guidance to public administrations on AI and language technologies.

Cultural Program on AI-based Language Technologies

- Implements a cultural program centered around AI for languages.
- Aims to engage the public in discussions and awareness about AI and language technologies.

Empowering Language Technologies End-Users

- Focuses on empowering Language Technology end-users.
- Encourages end-users to become contributors and active participants in the Language Technology landscape.

Multilingual Context

- Addresses challenges of AI and language technologies in a multilingual context

DIGITAL Work Programme 2024 (3 calls)

DIGITAL-2024-AI-06-LANGUAGE-01

Alliance for Language Technologies:

Data and Finetunig

€ 20 millions (Simple Grant)

- Support the Collection of Language Data

- Support Fine-tuning of Large Language Models

DIGITAL-2024-AI-06-LANGUAGE-02

Alliance for Language Technologies:

Ecosystem

€ 4 millions (CSA)

Support the Coordination of a European Language Technologies Ecosystem.

DIGITAL-2024-AI-06-FINETUNE

Open-source European foundation model

€ 25 millions (SME Grant)

Develop and make available one open-source large language foundation model as an infrastructure designed to be largely used by public or private users.

Thank you

European Commission