

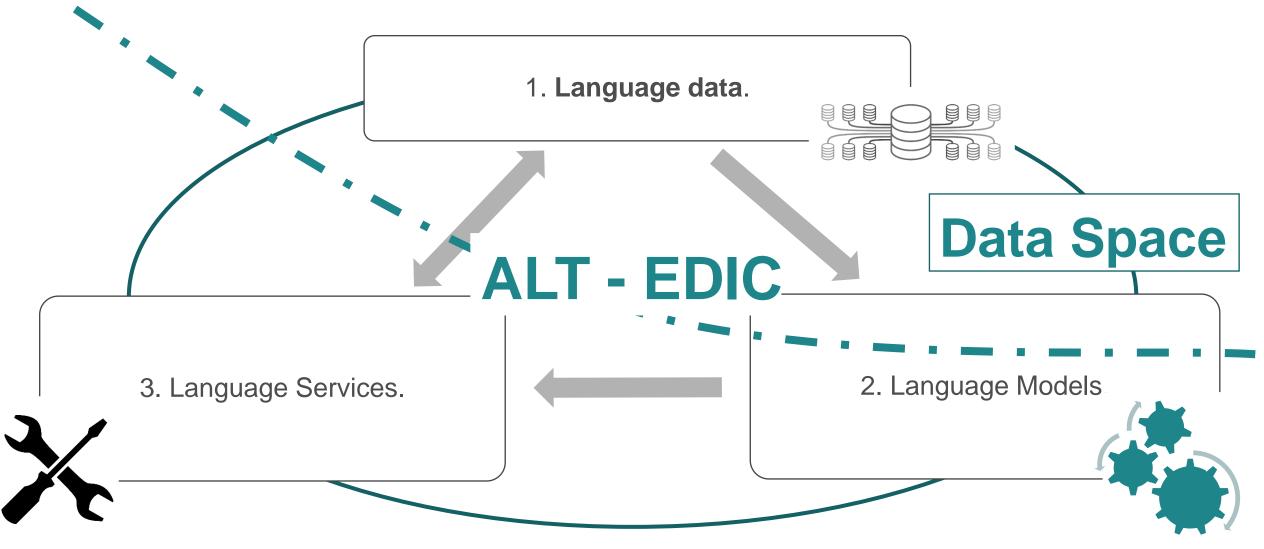
The Digital Europe Programme and the Common European Language Data Space

LDS workshop in Czechia

Language Technology for a Multilingual Europe

Philippe GELIN DG CONNECT – G3

European Language Technologies Landscape



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



- > ALT-EDIC
- DIGITAL calls 2024



The Common European Language Data Space (LDS)

- Overview
- Architecture



Language Data Space – Overview

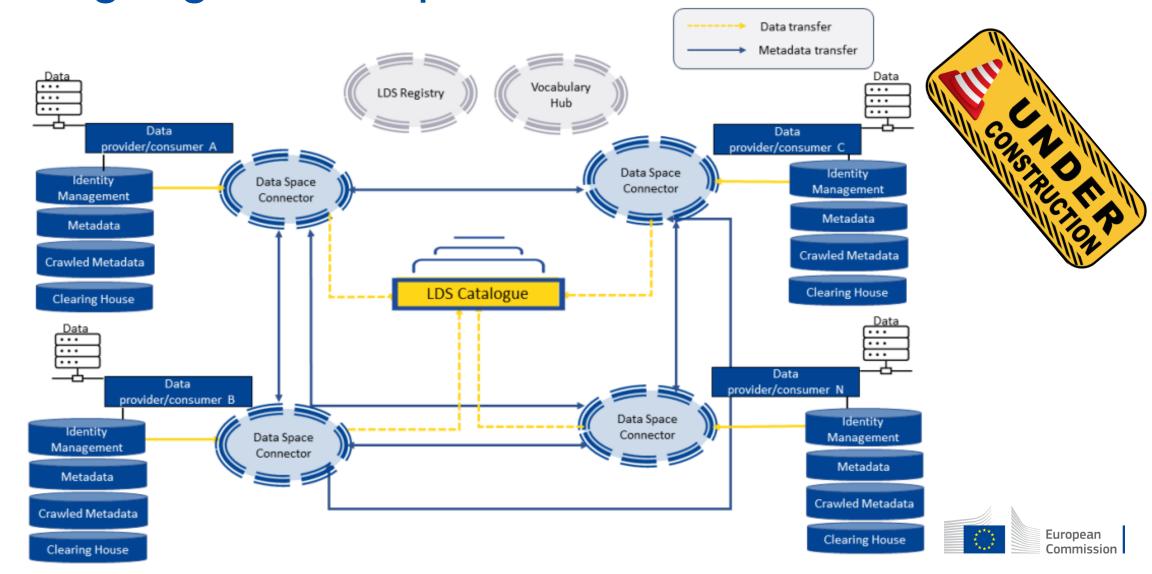


What	Platform and marketplace for the collection, creatio language data	n, sharing and re-use of multilingual and multimodal	
vviiai	6 Mio € Procurement Contract from the DIGITAL Work Programme 2021-2022		
\//b\/	Increase availability of clean, high-quality, compliant language data		
Why	Support the development of state-of-the-art European language technologies		
	Consortium: DFKI, ELDA, TILDE, Athena Institute		
	Governance: Member States representatives and Industry User group		
V V I I O	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		



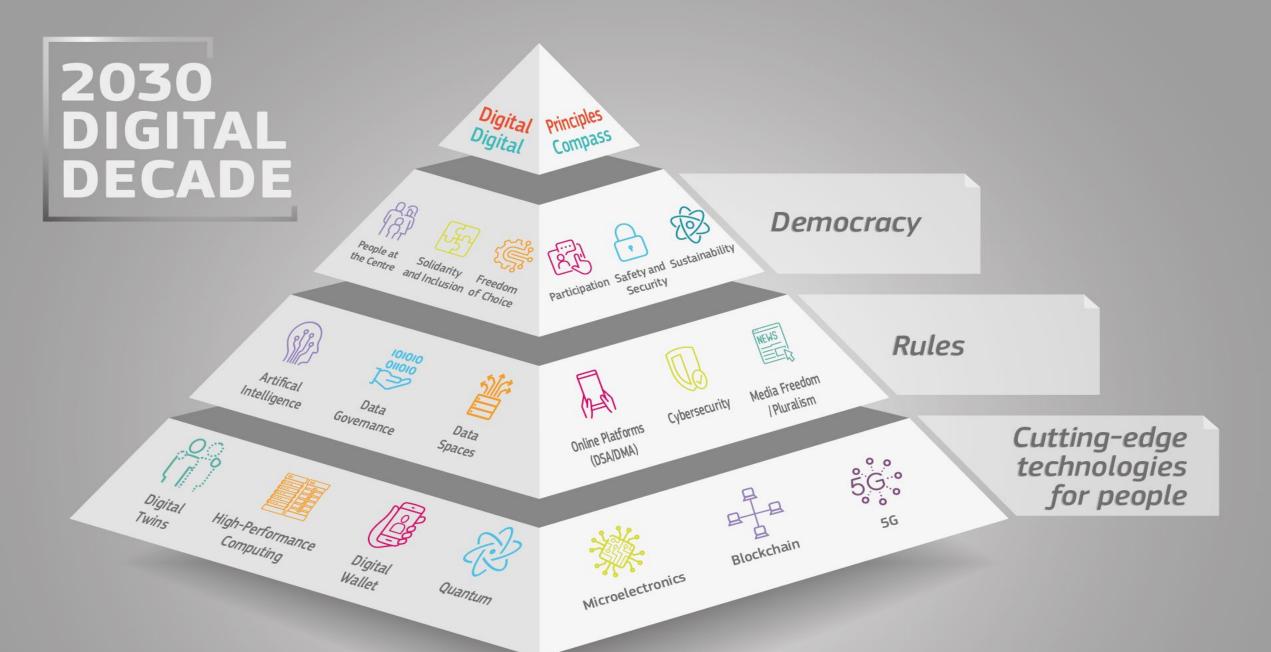
Language Data Space – Architecture





Alliance for Language Technology EDIC (ALT-EDIC)





ALT-EDIC – Alliance for Language Technologies <u>EDIC</u>

What	DDPP Multi-Country Project in the area of "European Common Data Infrastructure and Services"	
vviiat	Build a European ecosystem around Language Data and Technologies	
Why	Federate data and resources from EU and Member States to boost Data and AI competitiveness	
vviiy	Promote technology and service deployment for European leadership and strategic autonomy	
	Coordinated by: France (Ministry of Finance and Digitalization & Ministry of Culture)	
Who	Members: FR, ES - BG, CZ, DK, FI, GR, HR, HU, IE, IT, LT, LU, LV, NL, PL, SI	
	Observers: AT, BE, CY, EE, MT, PT, RO, SK – Interested: DE, SE	
Ном	~25 Mio € Member States contribution	
How	~25 Mio € EU funding (DIGITAL Programme)	
When	7 February 2024: Set up by Commission Implementing Decision (EU) 2024/458	
	Q1 2025: Fully operational	



ALT-EDIC

Alliance for Language Technologies European Digital Infrastructure Consortium

Objectives

Preserve linguistic and cultural diversity in Europe

Technological leadership and strategic autonomy

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]

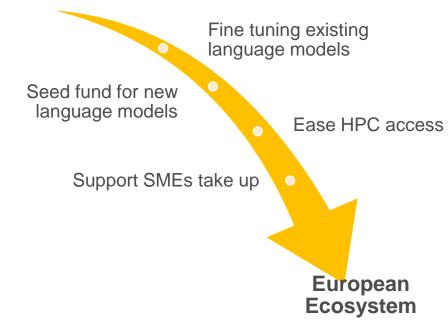
DE, SE

BUDGET

50M€

Actions Plan

Federate MS and EU language data resources





Action Plan - 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

Open-Source Repository

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

Existing Models

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

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.

New Models

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

Advisory Hub

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

Cultural Program on Al-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.

Ecosystem

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.



Landscape

LT Crucial for Preserving Identity and Competitiveness

European Commission's Actions

Maintaining languages, cultures, and diversity

Achieving efficiency in a multilingual Europe

Turning diversity into opportunities

GenAl4EU **AI Factories** Al Talent Compute Data **Data Strategy**

European

Thank you



© European Union 2023

Unless otherwise noted the reuse of this presentation is authorised under the <u>CC BY 4.0</u> license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Slide xx: element concerned, source: e.g. Fotolia.com; Slide xx: element concerned, source: e.g. iStock.com

