The Alliance for Language Technology EDIC and the Common European Language Data Space

LDS workshop in Estonia

Philippe Gelin
Head of Sector - DG CONNECT – G3

Language Technology Landscape

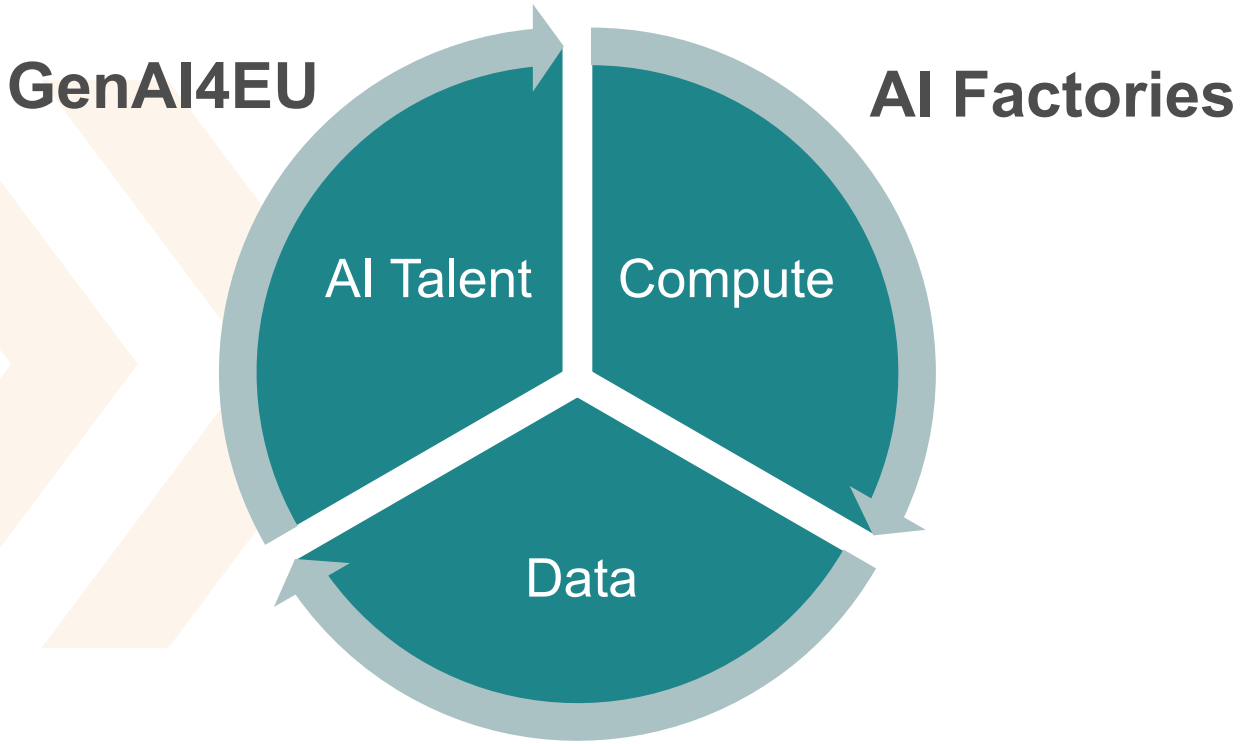
European Commission

Language Technology Landscape

LT Crucial for Preserving Identity and Competitiveness

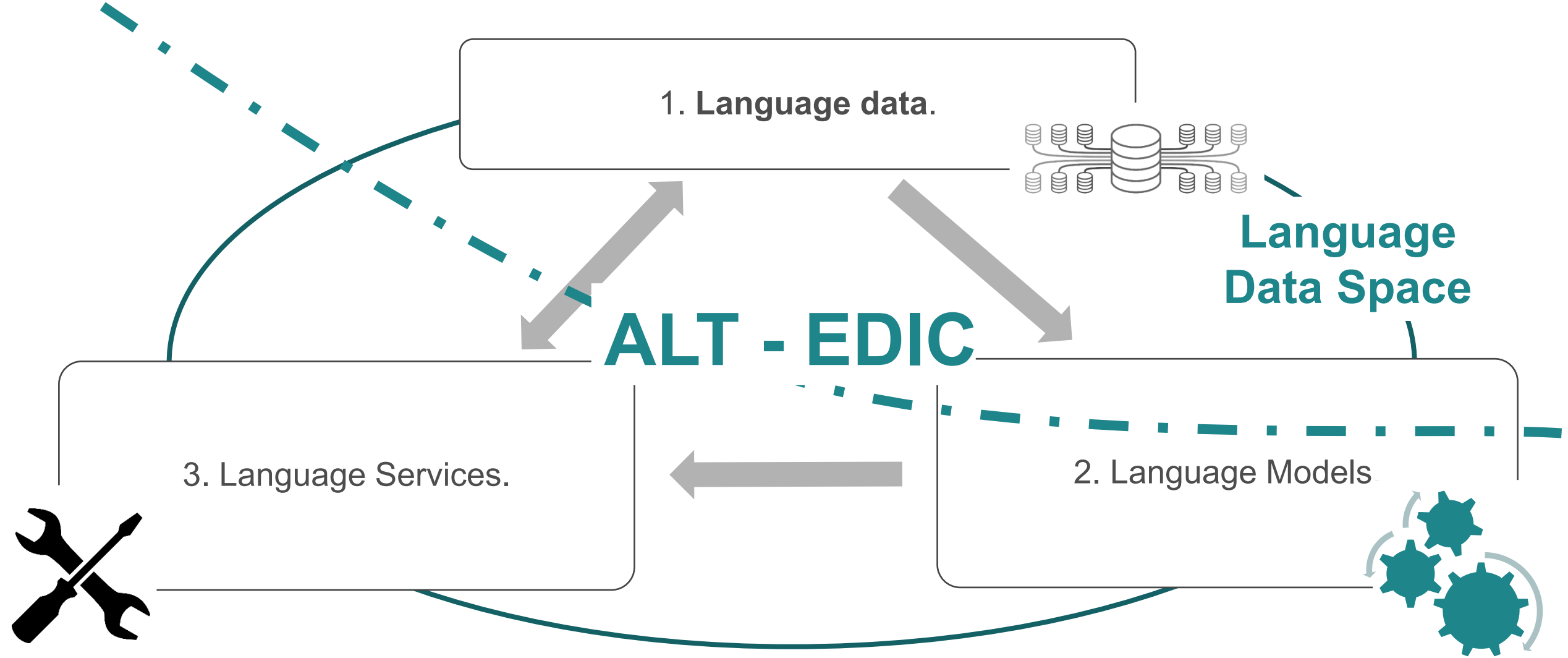
- Maintaining languages, cultures, and diversity
- Achieving efficiency in a multilingual Europe
- Turning diversity into opportunities

European Commission's Actions



European Commission

European Language Technologies Landscape



The Common European Language Data Space (LDS)

- Overview
- Architecture

European Commission

Language Data Space – Overview

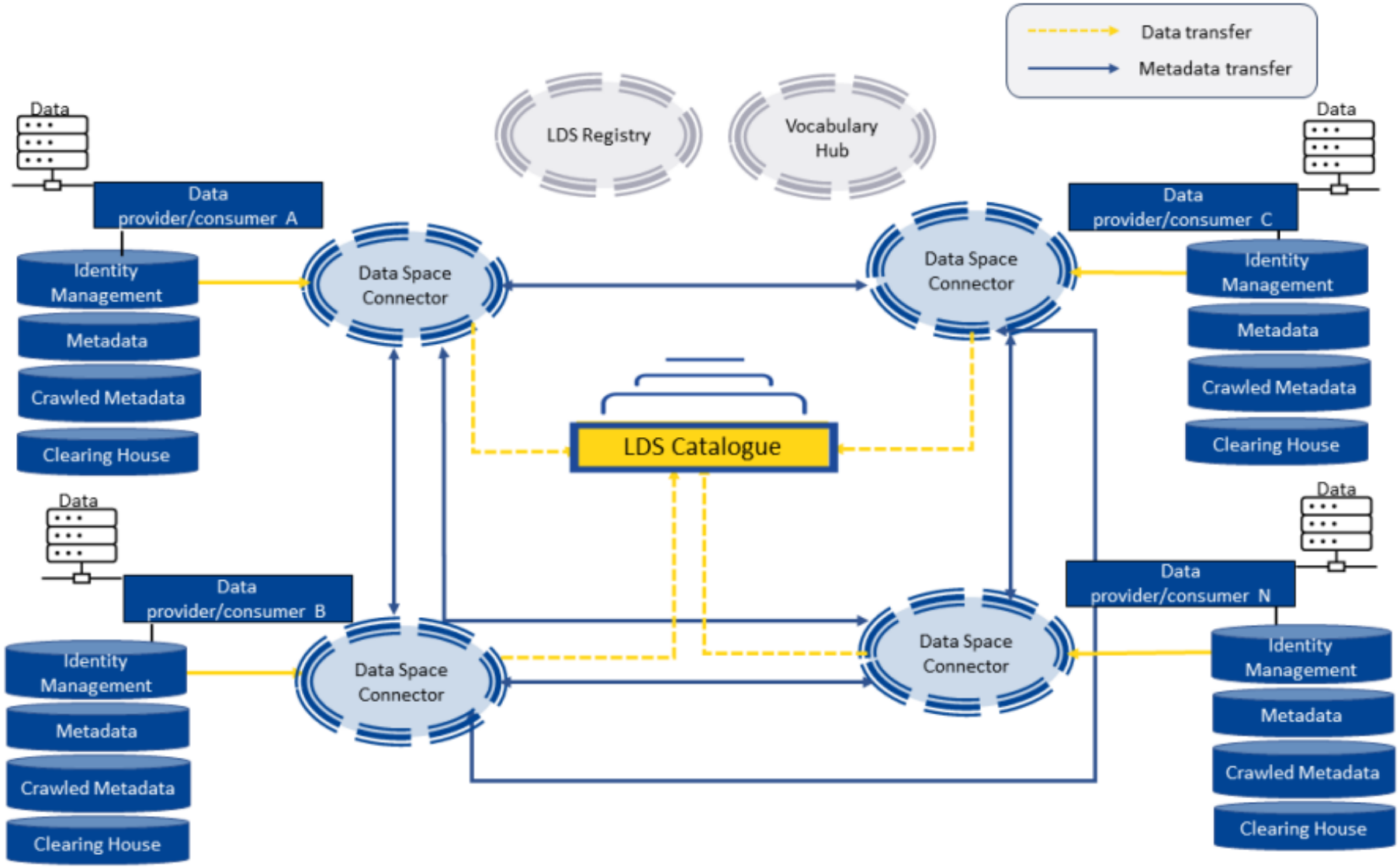
EUROPEAN LANGUAGE DATA SPACE

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	- Openness
	- 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).	
	Beta version in **Q1 2025**	

European Commission

Language Data Space – Architecture

EUROPEAN LANGUAGE DATA SPACE

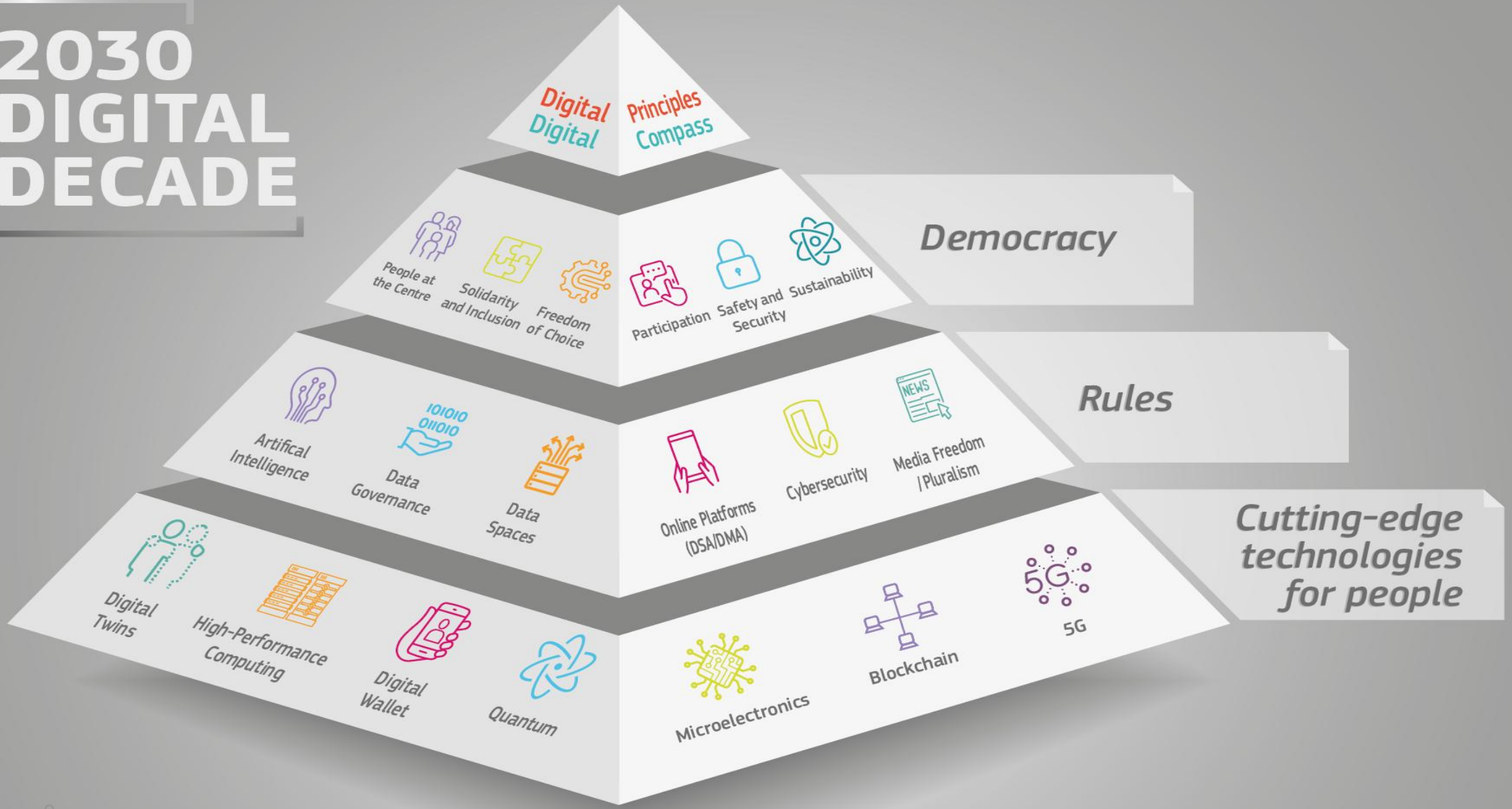


European Commission

Alliance for Language Technology EDIC (ALT-EDIC)

European Commission

2030 DIGITAL DECADE



#DigitalEU

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 [18]

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

Observers [8]

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

Interested [2]

DE, SE

Budget

Projects (Costs)	
- LLMs4EU	€ 40 Mio
- ALT-EDIC4EU	€ 4 Mio
- OpenEuroLLM	€ 40 Mio
- LLM-BRIDGE	€ 6,8 Mio

Actions Plan

Federate MS and EU language data resources

Seed fund for new language models

Fine-tuning existing language models

Ease HPC access

Support SMEs take up

European Ecosystem

European Commission

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

European Commission

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

European Commission

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

European Commission

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

European Commission

DIGITAL Work Programme 2024 (3 projects)

ALT-EDIC4EU

€ 4 millions
Coordination and Support Action (100%)

Support the **Coordination** of a European Language Technologies Ecosystem

LLMs4EU

€ 20 millions
Simple Grant (50%)

Support the Collection of Language Data

Support Fine-tuning of Large Language Models

OPENEuroLLM

€ 20 millions
SME Grant (75%)

Develop and make available one open-source large language foundation model

European Commission

One Last Thought: Why it matters?

- Join forces with 18+ Member States to co-develop models and services
- Support Local Economy in adopting multilingual AI solutions
- Ensure your language is represented in European AI and language models
- It’s not too late— You can still join ALT-EDIC as a full member
- Benefit from FSTP (cascade funding) opportunities in ongoing Digital Europe projects

Thank you

© European Union 2023

Unless otherwise noted the reuse of this presentation is authorised under the CC BY 4.0 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

European Commission