The European Language Data Space

—

Philippe Gelin

Head of Sector Multilingualism,
Unit "Accessibility, Multilingualism & Safer Internet"
Directorate Data
Directorate-General for Communications Networks, Content and Technology
European Commission

MY EUROPE, MY LANGUAGE.

...MADE IN EUROPE

Digital Programme

The [Digital Europe Programme (DIGITAL)](#) provides funding for projects in five crucial areas:

SO1

HIGH PERFORMANCE COMPUTING

SO3

CYBERSECURITY

SO4

ADVANCED DIGITAL SKILLS

DIGITAL EUROPE PROGRAMME

SO2

CLOUD, DATA AND AI

SO5

BEST USE OF DIGITAL TECHNOLOGIES

Programme *VISION*
Digital Decade and digital targets for 2030.

Programme *FOCUS*
Building the **strategic digital capacities** of the Union and facilitating the **wide deployment of digital technologies**.

Programme *GOAL*
Bridging the **gap between digital technology research** and **market deployment**, while supporting the EU **twin objectives** of a green transition and digital transformation.

DSs and the LDS are under SO2 'Cloud, Data and AI', whose aim is to drive the *digital transformation* of the *public and private sectors*.

European Commission

'Origins' of the Common European Data Spaces

Common European Data Spaces   are the centrepiece of the   **European Strategy for Data**

They are meant to harness the **value of data** for the benefit of the European economy and society.

Objectives

In February 2020, it announced the Commission's vision to empower businesses, public administrations and individuals to make better decisions **through the use of data**.

The creation of EU-wide common, interoperable DSs in strategic sectors is to **overcome existing legal and technical barriers** to data sharing and, as such, **unleash the potential of data-driven innovation**.

→ More data shared in a privacy-compliant, trustworthy, and secure way.
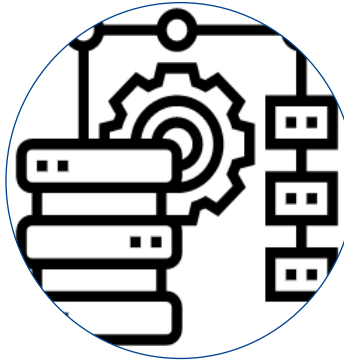
→ A genuine single market for data.

European Commission

Design Principles of the DS
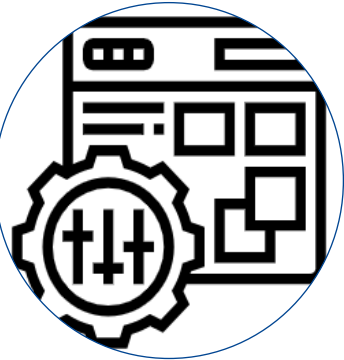
Data control

For data holders:
- Full control over data at all times;
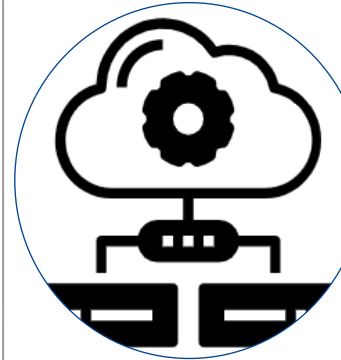- Option to change data access rights over time.

Governance Framework

- Fair,
- transparent,
- proportionate,
- non-discriminatory access to & sharing/use of data.

Respect of EU Rules & Values

- Personal data protection
- Consumer protection legislation
- Competition law

Technical Data Infrastructure

- Building blocks
- Efficient & coordinated DSs construction
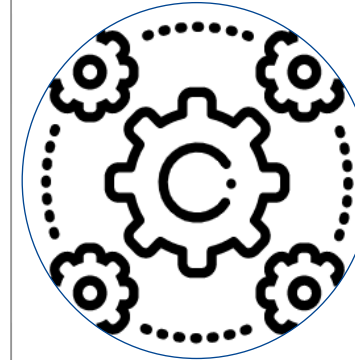
Interconnection & Interoperability

- Progressive interlinking of DSs

Aims:
- Fully integrated ecosystem
- Cross-exploitation of data

Openness

- For all actors, organisations or individuals;

Aims:
- Fostering competition
- Avoiding lock-ins
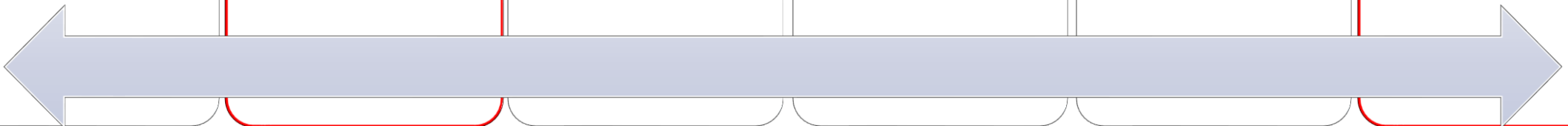
DIGITAL WP 21-22 Data Spaces at a glance

Green Deal	Smart Communities	Mobility
Manufacturing	Agriculture	Cultural Heritage
Health, Genomics	Health, Cancer Images	Media
Financial	Skills	Language
Public Procurement	Security and Law Enforcement	Tourism

European Commission
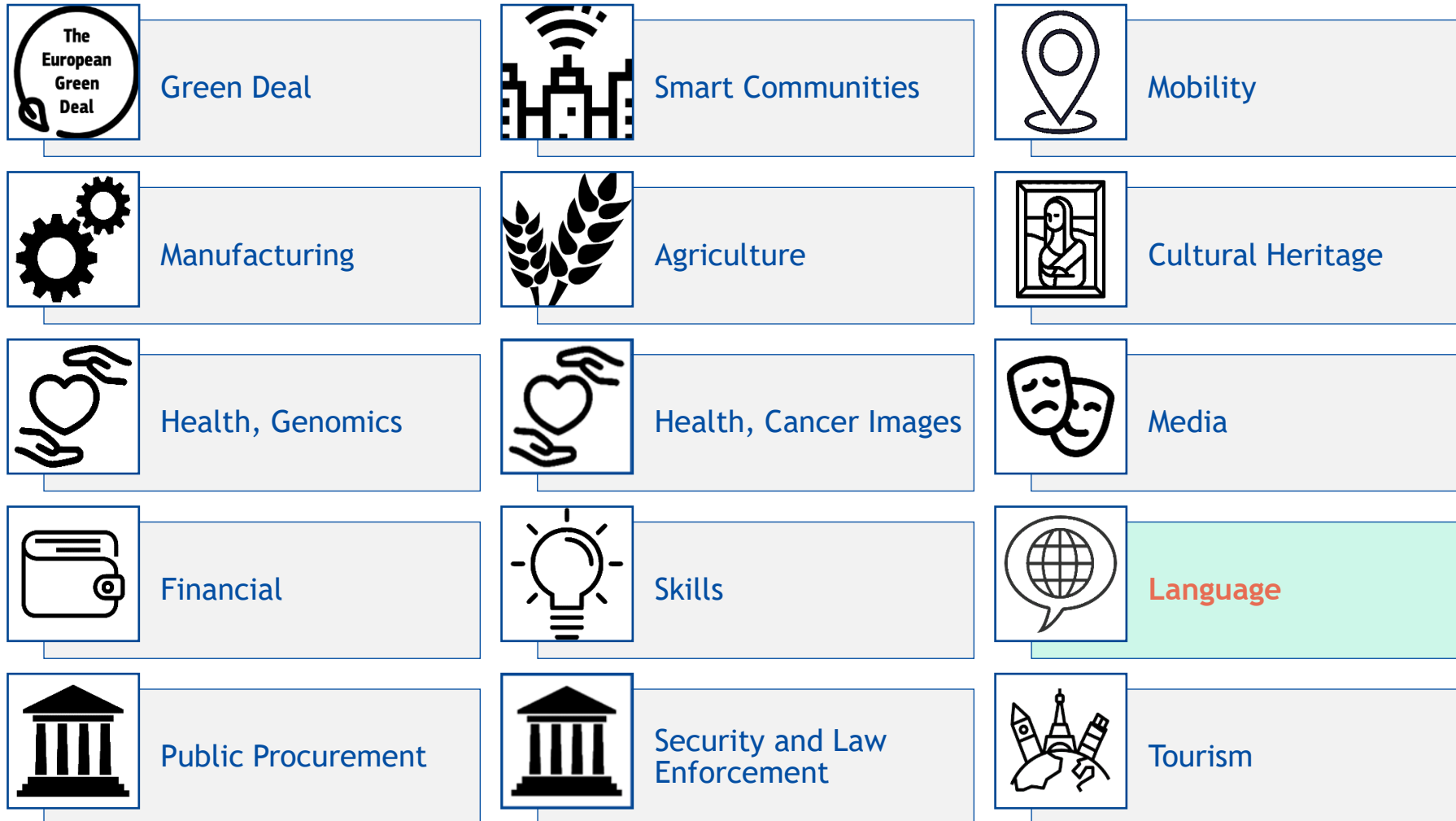
Language Data Space… What is it?

Increase Data Exchange

Goal

Where we are

Text, audio and video files are a large part of the data we produce daily.

To process and extract relevant information from these data, we need to have complex AI-based language services available.

Complex AI-based language services need large sets of data to be trained.

What we offer

Multimodal language data

Language Data Space

| Collection | Creation | Sharing | Re-use |
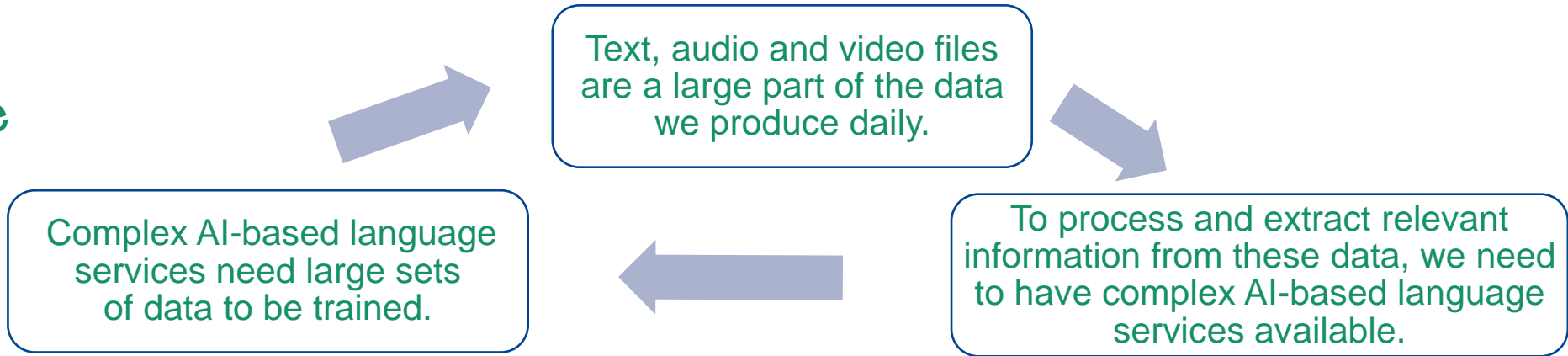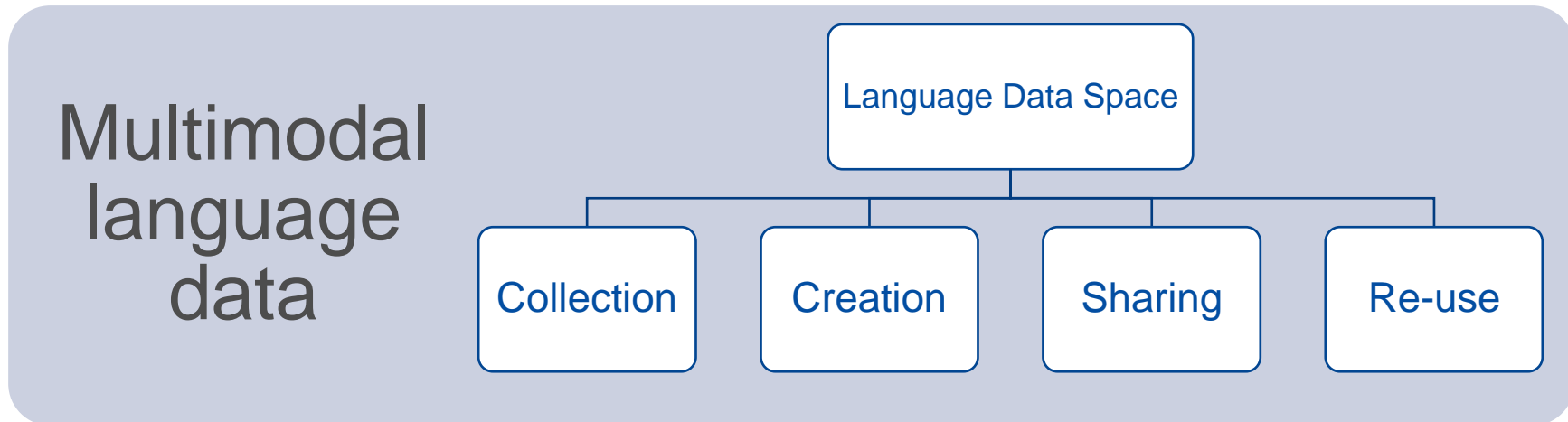
European DS Ecosystem – DIGITAL SO2

EDBI - European Data Innovation Board
supporting the EC in issuing guidelines on how to facilitate the development of DSs as well as identifying the relevant standards and interoperability requirements for cross-sector data sharing.

SO1
HPC

SO2
Cloud, AI, Data

SO4
Advanced Skills

SO3
Cyber-security

SO5
Best use of Tech
Common Services Platform

European Data Spaces Technical Framework

consisting of

Cloud-to-Edge Infrastructures and Services

including

- Smart Middleware Solutions
- Marketplace

DSSC

Sectoral Data Spaces

Media   Manufacturing   Agriculture

Smart Cities   Mobility   …

LANGUAGE DATA SPACE

DSSC – Data Spaces Support Centre
support all relevant actions on sectoral DSs and their architectures and data infrastructure requirements available for the DSs.

AI on-demand platform
connecting providers and consumers/users of AI resources.

TEFs – Testing and Experimentation Facilities
for experimenting with AI-based solutions in real-world environments.

LDS Work Strands

1 **Establishing** an institutional Centre of Excellence for Language Technologies (**CELT**)

2 **Deploying** the Language Data Space

Actions:

Governance layer
→ to framework the collection, creation, sharing and re-use of **language data & models**, incl. legal aspects.

Operational layer
→ to implement the **architecture**;
→ to promote the **exchange of data**;
→ to support the deployment of **large multimodal language models** and a wide range of **AI language technology services** through the AI on-demand platform.

Establishment of the CELT & the CELT+

CELT Coordinating across the MSs the creation and collection of multimodal language data and models.	*Networking*	Identifying and bringing together existing European **stakeholders/initiatives** and **language resources** • Datasets and data streams • _Large language models_ *ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, META-SHARE, etc.*
	Governance	Developing a **Multi-Stakeholder Data and Services Governance Scheme**
CELT+ Driven by Industry	*Planning*	Elaborating a **Blueprint for the Language Data Ecosystem** based on existing EU/national legislation, incl. a detailed **roadmap** on how to deploy the language data space

Strategic Level (left axis, upper)
Operational level (left axis, lower)

EC Role: Facilitator

European Commission

LDS Distributed Architecture
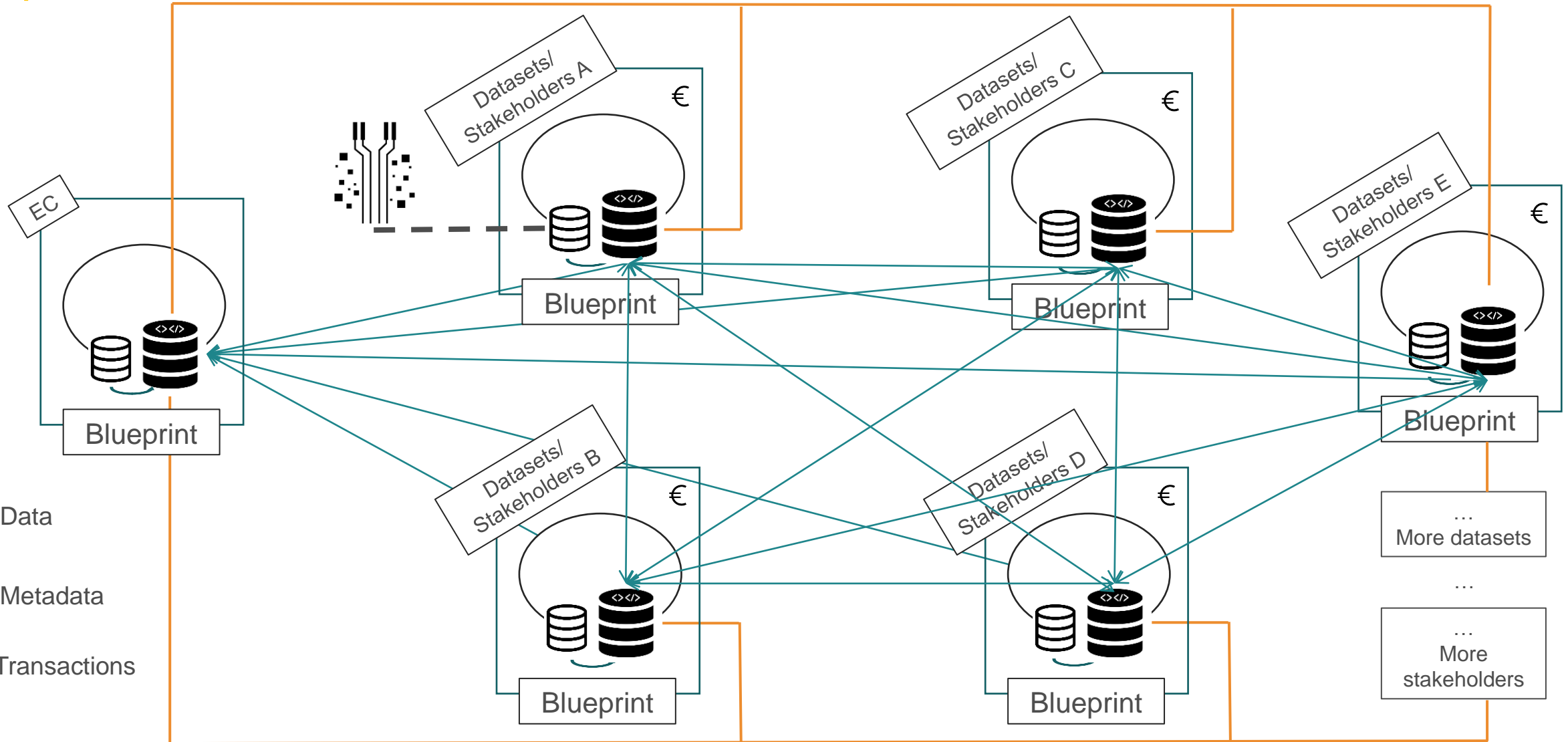
Stakeholders' Landscape

Other Data Spaces

Smart Middleware
= cloud-to-edge
infrastructure

Governance
of the distributed
data
Infrastructure

PRIVATE

PUBLIC

Industry
- Text
- Audio
- Video
- (Language) **Models**

Media
- Text
- Audio
- Video
- **(Media) Streams**

Research Institutes
- Text
- Audio
- Video
- (Language) Models

Cultural Associations
- Text
- Audio
- Video

Public Administrations
- Text
- Audio
- Video

Business Intelligence
Access to internal and
international market(s)
Customised services
Economic Recovery
Industry

Business Intelligence
Access to networks of
stakeholders and potential
customers
Personalised services
Media

Access to data | **Researchers**

Better Information
Transparency
Access to digital services
supported by language
technologies
Citizens

More and better
services offered
Digitisation
Public Administrations

AI on-Demand Platform

HPC, TEFs, EDIHs, …

ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, META-SHARE

European Commission
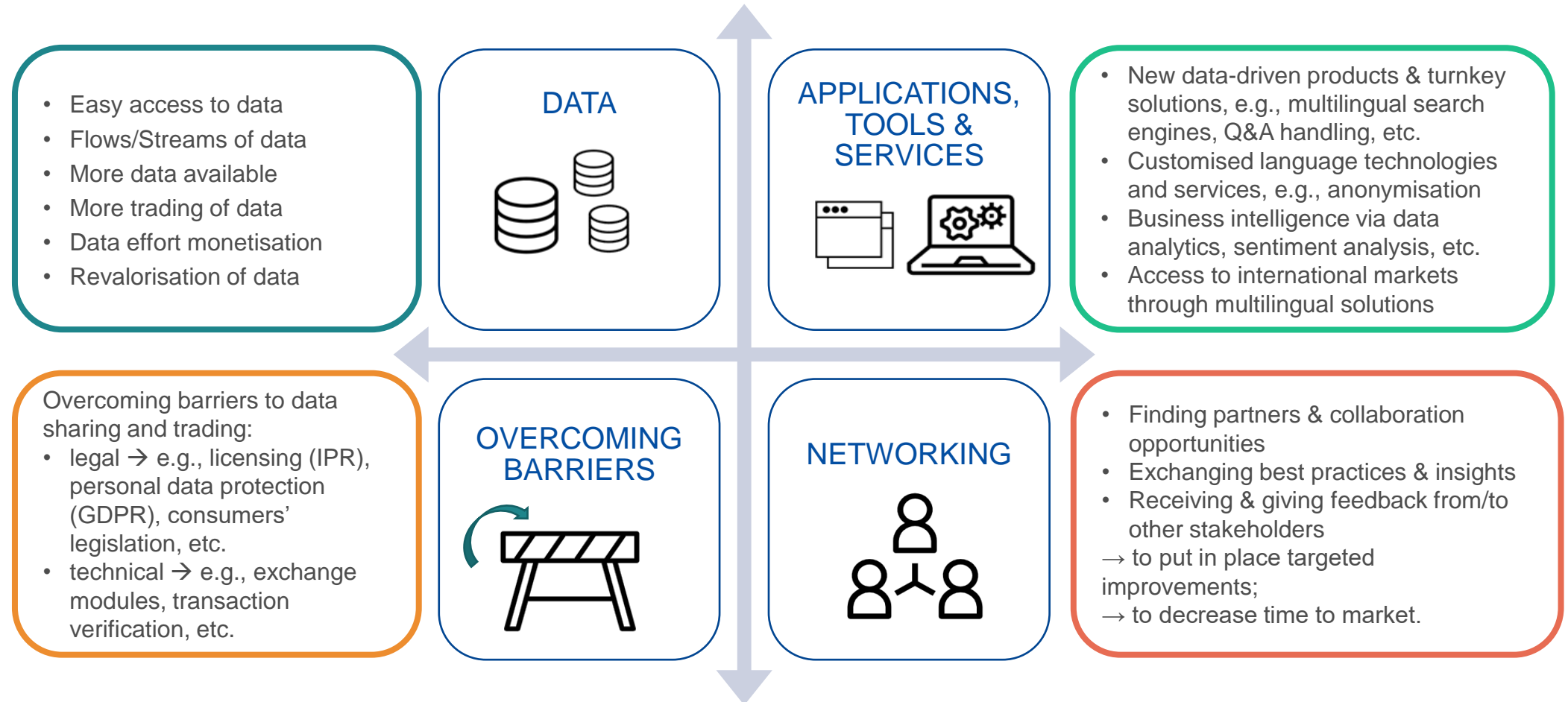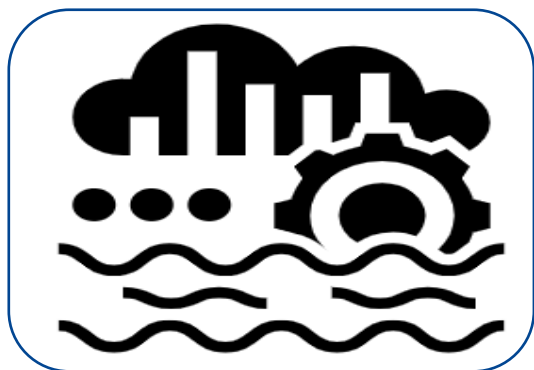
Benefits of the LDS

The LDS has been thought to be **a platform and a marketplace** *where the request for language resources and services matches the offer*.

- Easy access to data
- Flows/Streams of data
- More data available
- More trading of data
- Data effort monetisation
- Revalorisation of data

DATA

APPLICATIONS, TOOLS & SERVICES

- New data-driven products & turnkey solutions, e.g., multilingual search engines, Q&A handling, etc.
- Customised language technologies and services, e.g., anonymisation
- Business intelligence via data analytics, sentiment analysis, etc.
- Access to international markets through multilingual solutions

Overcoming barriers to data sharing and trading:
- legal → e.g., licensing (IPR), personal data protection (GDPR), consumers' legislation, etc.
- technical → e.g., exchange modules, transaction verification, etc.

OVERCOMING BARRIERS

NETWORKING

- Finding partners & collaboration opportunities
- Exchanging best practices & insights
- Receiving & giving feedback from/to other stakeholders
→ to put in place targeted improvements;
→ to decrease time to market.

Key messages

Data sharing against remuneration	Developing, **selling** and **purchasing** language-based applications, **tools** and **models**	Developing, **selling** and **purchasing** language-based services	Performing data analytics and further exploiting the market for more reach and visibility
E.g., multilingual and/or multimodal data	E.g., automatic speech transcription	E.g., automatic text summarisation	E.g., historical data

European Commission

Previous Projects Relevant for the LDS

ELRC

European Language Resource Coordination

under CEF - 'Connecting Europe Facility'

ELG

European Language Grid

under HORIZON 2020

Language Data Space

It has focused on collecting open language resources and language processing services, defining collection guidelines and promoting the ELRC-SHARE, a repository used for documenting, storing, browsing and accessing language data and tools.

It develops and deploys a scalable cloud platform, providing access to hundreds of commercial and non-commercial Language Technologies for all European languages, including running tools, services, data sets and resources.

European Commission

Our actions: Legislative, Coordination & Funding

Next Generation EU & Recovery and Resilence Facility

Next Generation EU #NextGenerationEU #EUBudget

40 % 26 %

Green transition — Digital transformation — Smart, sustainable and inclusive growth

Social & territorial cohesion — Health, and economic, social and institutional resilience — Policies for next generation
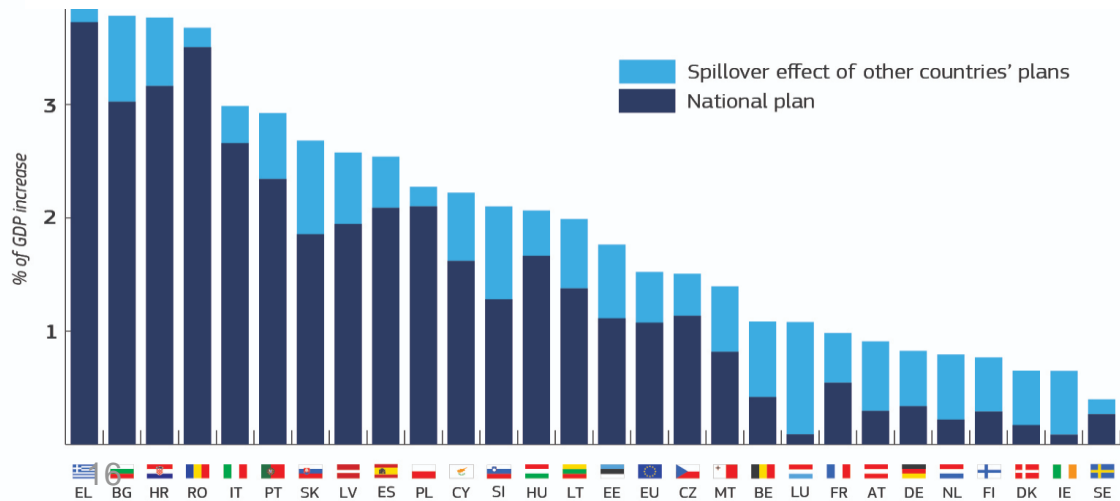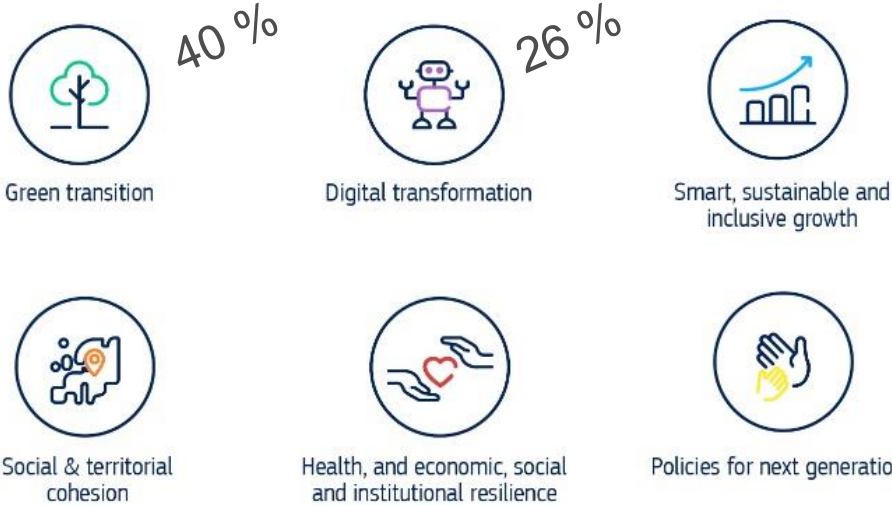
Recovery and Resilience Facility (RRF)	€723.8 billion
of which, loans	€385.8 billion
of which, grants	€338.0 billion

Spillover effect of other countries' plans
National plan

% of GDP increase

EL BG HR RO IT PT SK LV ES PL CY SI HU LT EE EU CZ MT BE LU FR AT DE NL FI DK IE SE

② Implementation until end-2026

Up to twice a year

MEMBER STATE Member States implement the reforms and investments set out in their plans

MEMBER STATE Upon completion of agreed milestones and targets, they request disbursements.

COMMISSION Then disburses the amount using funds raised by issuing bonds on behalf of the EU

COMMISSION The Commission as-sesses the request

After consulting expert committee

Within two months

FUNDS Funds are raised by issuing bonds on behalf of the EU

https://ec.europa.eu/info/business-economy-euro/recovery-coronavirus/recovery-and-resilience-facility_en

What is a European Digital Infrastructure Consortium ?

Pending Legal Implementation

- EDIC is a new legal framework to support Multi Country Project

- They have first been introduced in the Path to the Digital Decade Policy Programme (DDPP) proposal.

- The concept has been inspired by the success of European Research Infrastructure Consortia (ERICs).

- Advantages:

  ✓ Facilitating coordination of funding between MS;

  ✓ Speedy establishment and flexible implementation;

  ✓ Ensuring common standards and interoperability;

  ✓ Sustainability.

European Commission

EDIC vs Other MCP Implementation Mechanisms

Key Characteristics / MCP Implementation	Legal personality (mandatory)	Rapid set up	Outside the research realm	Membership: MS participation is mandatory	Membership: Flexibility (MS can join later)	Membership: Other entities than MS allowed (companies or intl organisations)	Governance: Voting rights proportionate to financial contributions of all entities	Financial liability: Limited to respective contributions of members	State Aid Rules: Assumption of Conformity	Exemptions from VAT and excise duty	Statutory Seat in one MS
Joint Undertaking (JU)	X	-	-	-	X	X	-	X	X	X	N/A
European Research Infrastructure Consortium (ERIC)	X	X	-	X	X	-	X	X	X	X	X
Through Union's agencies	N/A	-	X	N/A	N/A	N/A	N/A	N/A	X	X	N/A
Important Projects of Common European Interest (IPCEI)	-	-	X	X (Min. 2, after rev. will be 4)	-	X	-	-	X	-	-
European Digital Infrastructure Consortium (EDIC)	X	X	X	X (Min. 3)	X	X	X	X	X	X	X

What are the main features of an EDIC?

LEGAL STATUS
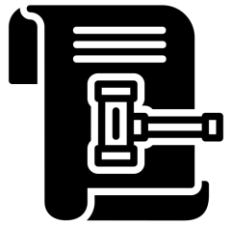
- Legal entity
- Set up by EC decision upon MS request
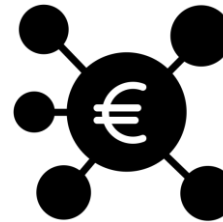
PARTICIPATION

- 3 MS needed for set-up
- Members = voting rights
- Observers = no voting rights
- Open for new MS to join
- Open for entities other than MS

GOVERNANCE

- Flexible structure, defined in the EDIC Statutes
- Assembly of Members with a Director

FINANCING

- Commitments of EDIC MS (financial or not)
- Funding from EU Programmes (CEF, DEP, HORIZON Europe, etc.)
- Contributions from private entities

FINANCIAL ADVANTAGE

- VAT exemption (if EDIC recognised as international body)

LIABILITY

- MS liability limited to their contribution/ (general)
- No EU liability for EDIC's debts

European Commission

What initiatives could a Language EDIC benefit from?

Pending Legal Implementation

From EC Side

- Ongoing **European Language Resource Coordination** (ELRC), finishing end of 2022

- Upcoming **LDS** procurement, starting in early 2023

- EU **research initiatives** (e.g., H2020 European Language Grid or ELG, ending in June 2022) Future HORIZON calls

From MS Side *(non limitative)*

- Germany: OpenGPT-X project

- France: BigScience project and 'Jean Zay' supercomputer

- Spain: PERTE – The New Language Economy

- Sweden: Autonomous Systems and Software Programme

- Slovenia: Development of Slovene in Digital Environment (DSDE)

- Latvia: Hugo.lv

- …

What are the next steps (1)?

EC side

- **EDIC** significant stages and next steps

  - Political agreement on the DDPP Decision reached in *July 2022*;
  - Informal Call for Expression of Interest in EDICs: *November/December 2022*;
  - Expected DDPP entry into force: *January 2023* (tbc).

  = Initiation of a Language EDIC as early as 2023, but sync with LDS necessary.

- EC guidance on implementing best-suited mechanism for the MCP in question.

The European Language Data Space

Philippe Gelin

Head of Sector Multilingualism,
Unit "Accessibility, Multilingualism & Safer Internet"
Directorate Data
Directorate-General for Communications Networks, Content and Technology
European Commission

MY EUROPE, MY LANGUAGE.

...MADE IN EUROPE