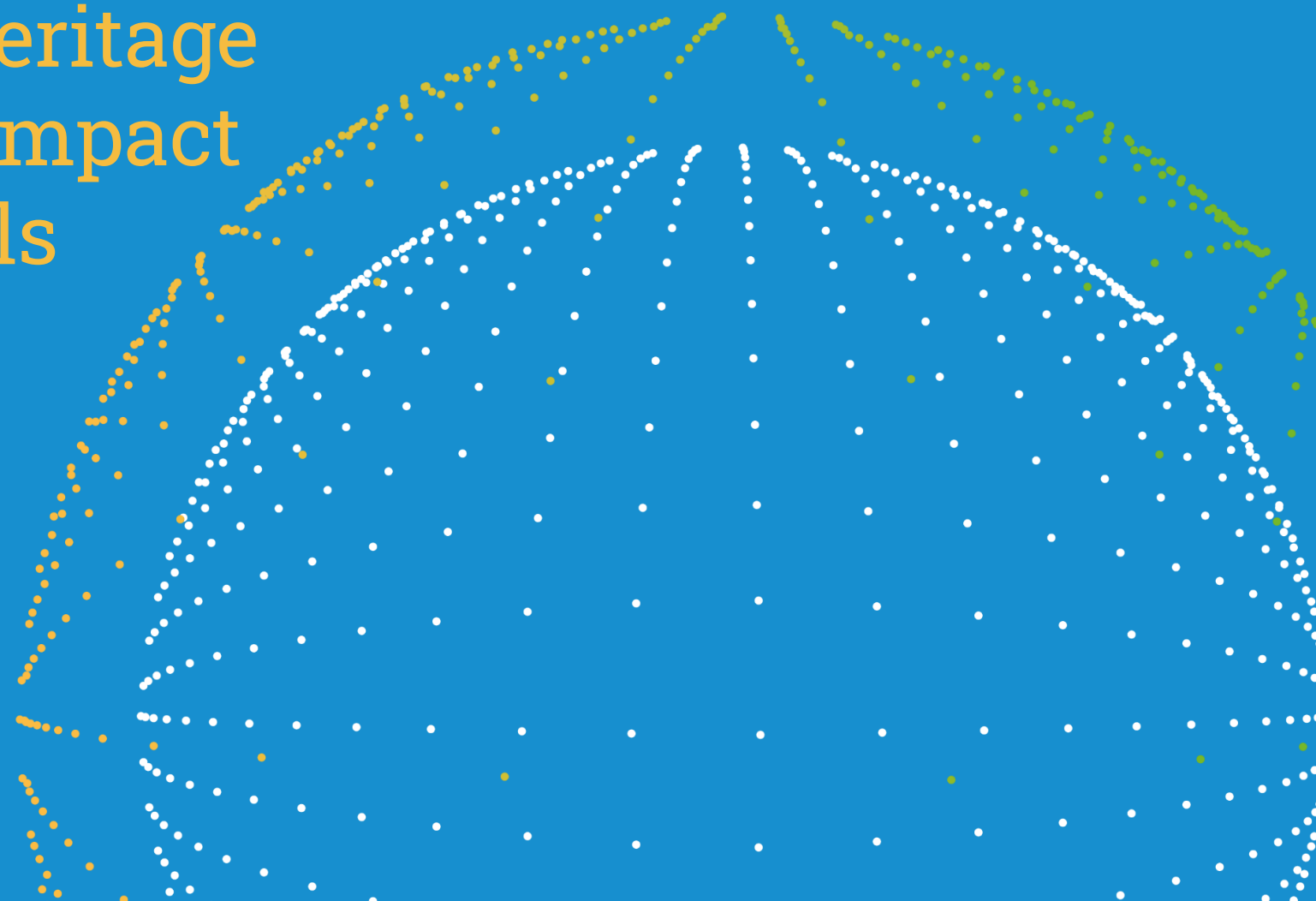


Data Spaces Symposium

Data, AI and beyond in
language, cultural heritage
and media sectors. Impact
on training and skills

Daniel Alonso, Georg Rehm,
Andrejs Vasiljevs, Oscar Rey,
Sylvain Le Bon, Valentine
Charles, Sabine Zander



Objectives of the session

- Identify **potential** of data and the **impact** of data sharing / data spaces on each specific sector
- Discuss about how to leverage **data and data sharing to power AI / Generative AI**, in those sectors
- Explore potential synergies and inter-sector connections, interoperability and potential joint use cases. Possible next steps

Data Spaces Symposium

Unite. Innovate. Adopt.

Data, AI and beyond in language, cultural heritage and media sectors

13 March 2024 | 15:45 - 16:55



Andrejs Vasiljevs
Tilde



Oscar Rey
Innovalia Association



Sylvain Le Bon
Startin'blox



Sabine Zander
imc



Valentine Charles
Europeana



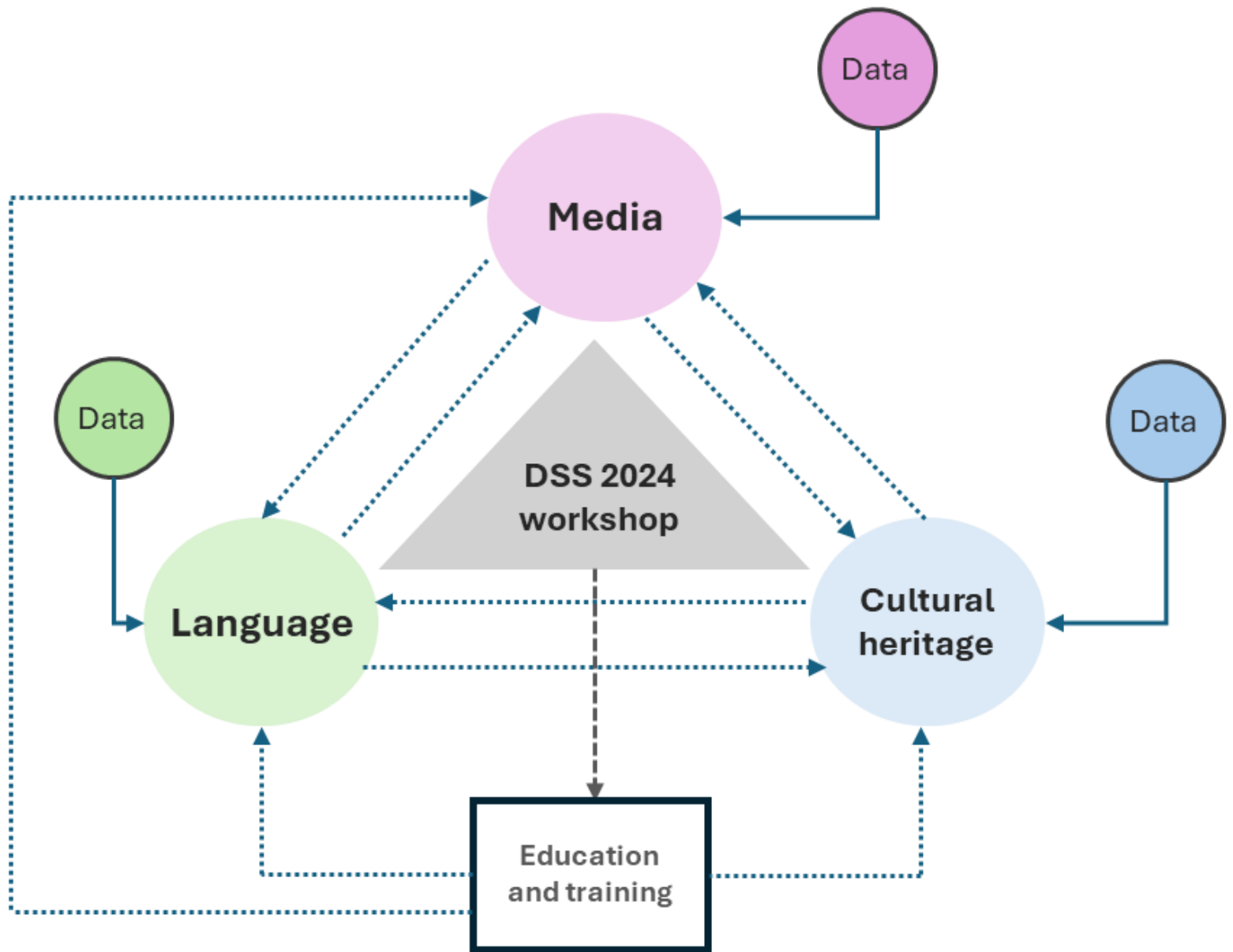
Georg Rehm
DFKI



Daniel Alonso
BDVA

Agenda

Introduction	
Daniel Alonso (BDVA)	Introduction and setting-up the scene
Initial statements and panel discussion	
Georg Rehm	View from language data space
Andrejs Vasiļjevs	Industry perspective, language and AI
Oscar Rey	View from media data space
Sylvain Le Bon	Industry perspective in the media sector
Valentine Charles	View from cultural heritage sector
Sabine Zander	Education benefiting from synergies. Relevance of up- and reskilling within these sectors (AI-assisted skill management)





EUROPEAN LANGUAGE DATA SPACE



European Language Data Space

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
georg.rehm@dfki.de

13-03-2024 Data Spaces Symposium, Session: Data, AI and beyond in language, cultural heritage and media
<https://language-data-space.ec.europa.eu>

Context: Large Language Models (LLMs)

- Large language models are the most disruptive breakthrough in AI in recent history (BERT, GPT-3, ChatGPT, GPT-4 etc.)
- LLMs are trained on vast amounts of training data (language data)
- LLMs use dozens, some even hundreds of terabytes (trillions of tokens) of language and also image, video, audio etc. training data
- Europe's languages are vastly under-resourced, except English, i.e., for many languages we have a substantial lack of data, which negatively impacts the performance of LLMs for those languages
- At the same time, the global LT/NLP Market is exploding: 439.85B\$ by 2030
- A concerted effort for the collection of enormous amounts of language data for all European languages is very much needed – we need new data, fresh data, industry data to compete.

Common 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 European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- Salient features: governance framework, technical architecture and infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens
- LDS is one of the 14 official EU data space projects – focus on industry

Consortium and Subcontractors

Lead Partner and Coordinator		
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	DE
Partners and Operation Leads		
R.C. "Athena", Institute for Language and Speech Processing	ILSP	GR
Evaluations and Language Resources Distribution Agency	ELDA	FR
TILDE	TILDE	LV
Main Subcontractors		
3pc GmbH Neue Kommunikation	3pc	DE
Capgemini Deutschland GmbH	CapG	DE
CLARIN ERIC	CLARIN	NL
Big Data Value Association (Data, AI and Robotics) AISBL	BDVA	BE

Plus legal experts (Delcade, France) and approx. 30 organisations
for the logistics of multiple country workshops

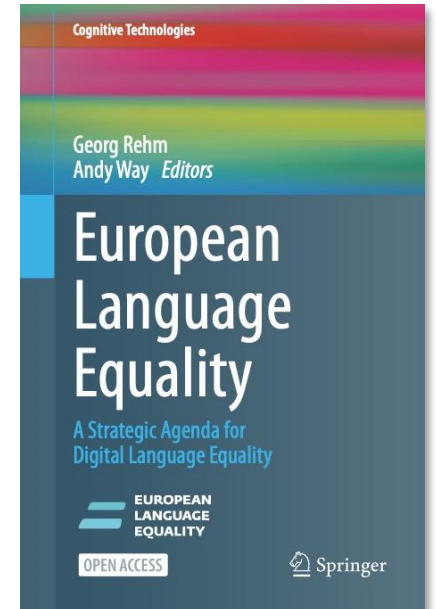
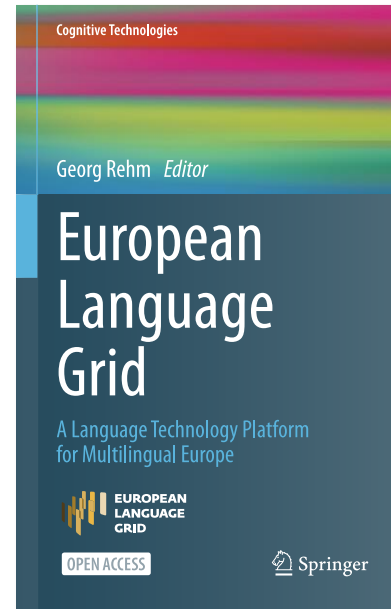
Previous Projects and Initiatives

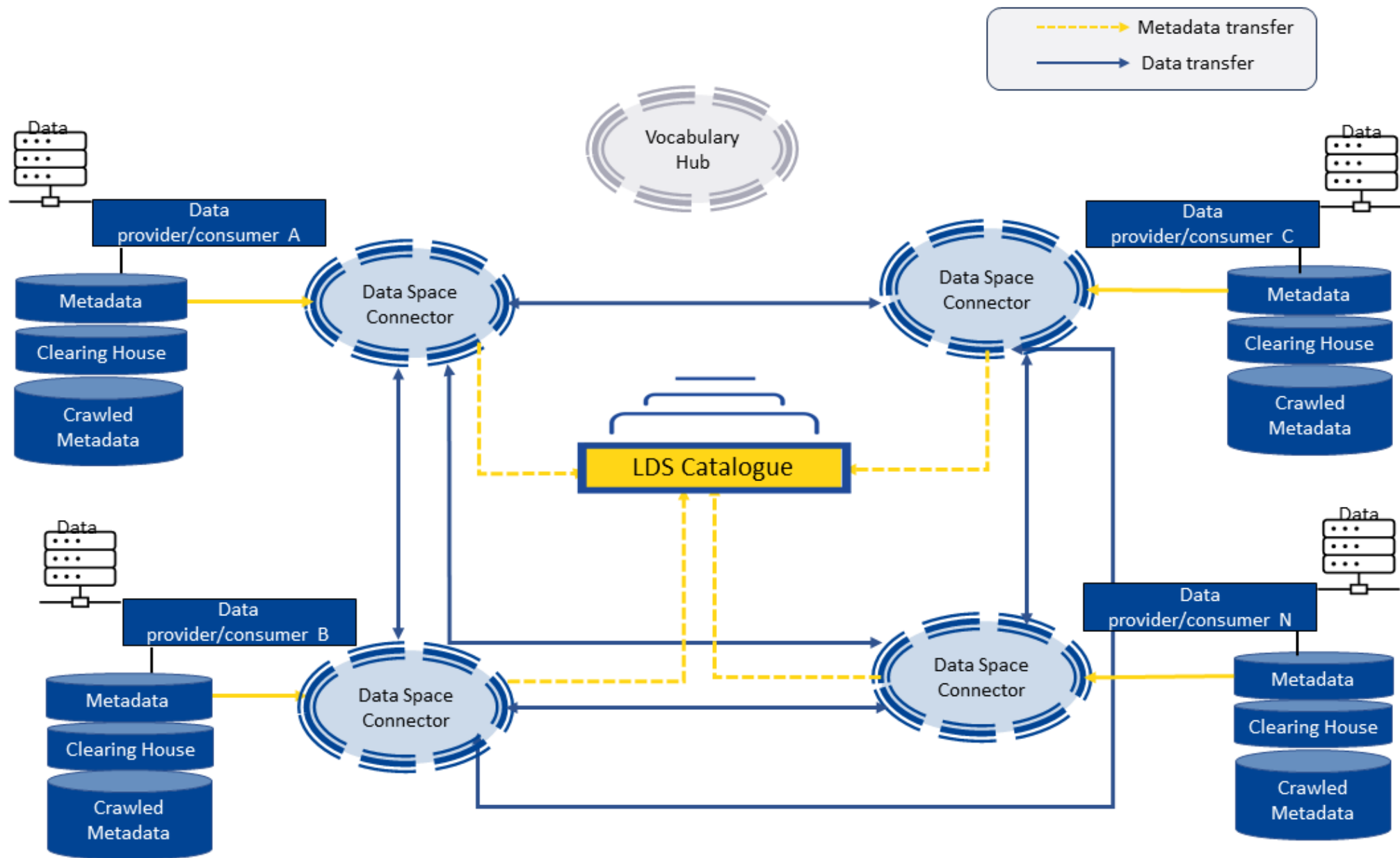
- The four core partners – DFKI, ILSP, ELDA, TILDE – have been involved in many projects, including:
- **META-NET** (FP7, 2010-2013)
 - META-SHARE
- **ELRC** (CEF, 2014-2023)
 - ELRC-SHARE
- **ELG** (H2020, 2019-2022)
 - ELG Cloud Platform
- **ELE** (PP/PA, 2021-2023)

META-NET



The **technical development work in LDS** will be informed by ELG, ELRC-SHARE, META-SHARE.





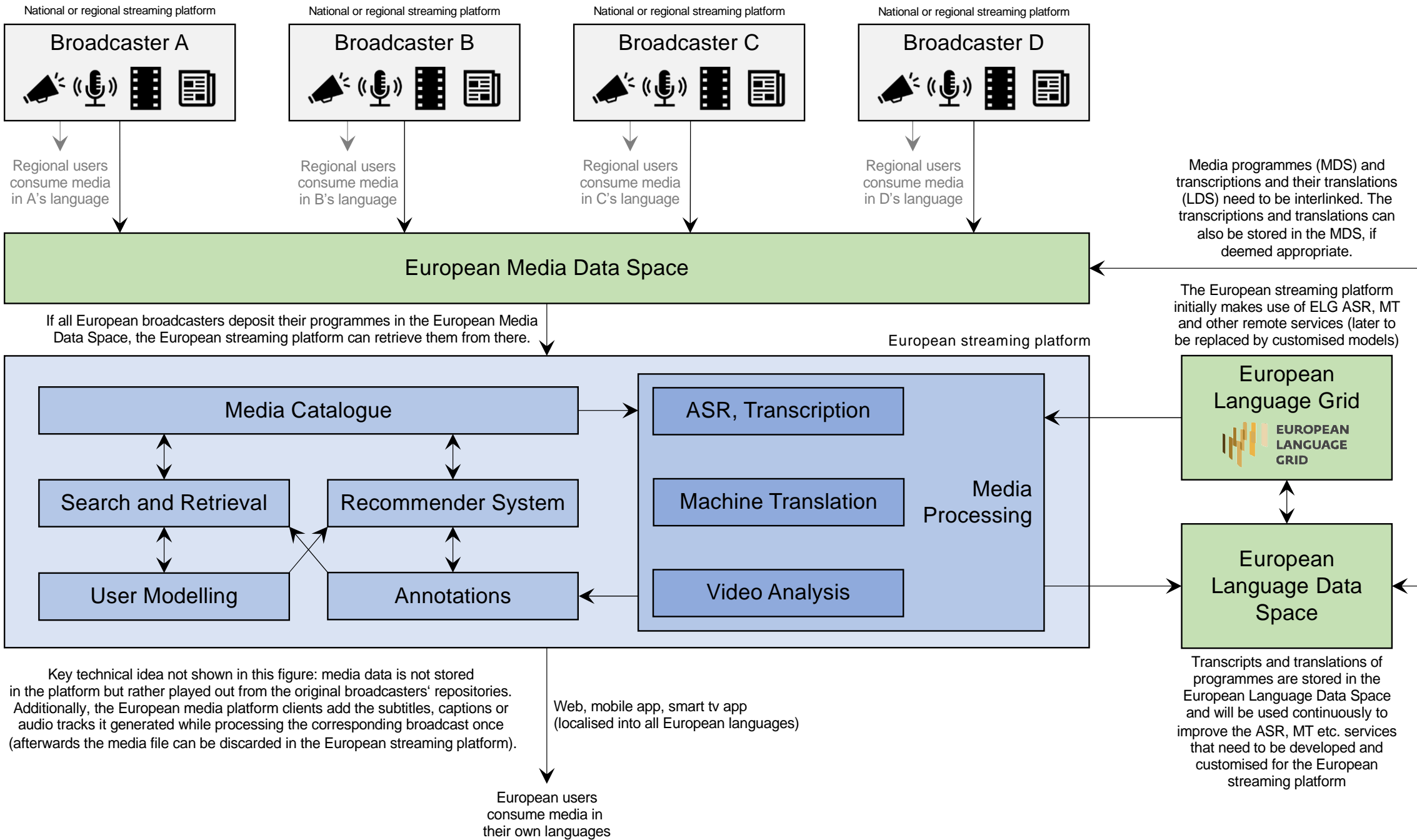
Classes of Data

Class of Data	Typical Size	Providers	Integration into LDS	Relevance for LLMs
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	Especially high quality data or domain-specific data or data covering specific languages and thus highly relevant for LLMs

Next Steps

- LDS is in full swing: technical development, promotion, dissemination, governance etc.
 - Collaboration with DSSC and ALT-EDIC
 - Collaboration with European projects, e.g., HPLT, OpenGPT-X, OpenWebSearch
 - Collaboration with data spaces, especially Media and Cultural Heritage
 - Collaboration with EuroHPC
 - Very important next step:
 - Adoption of LDS by industry and other organisations
 - Identify and make available new and fresh language data, especially from industry and covering all European languages and modalities
 - Most important synergy between LDS and other data spaces: new and fresh language data, especially media and broadcast data

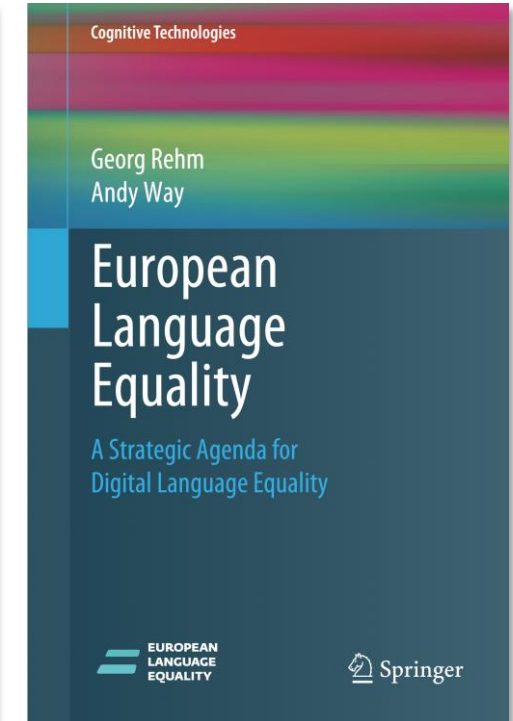
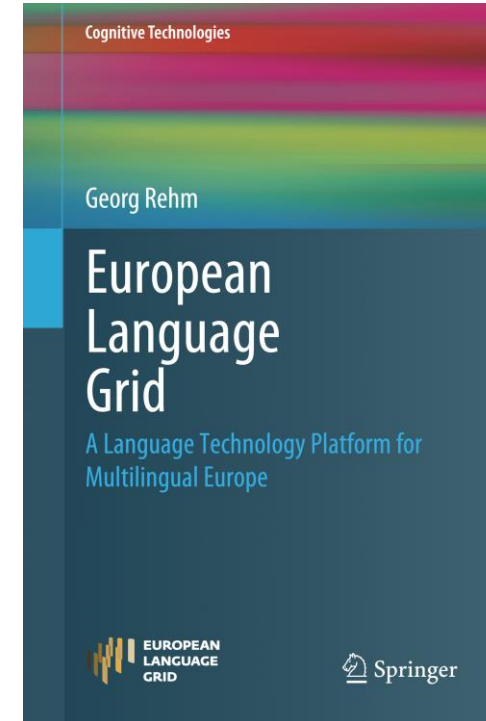






Common European Language Data Space

**Thank
you!**



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

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
georg.rehm@dfki.de

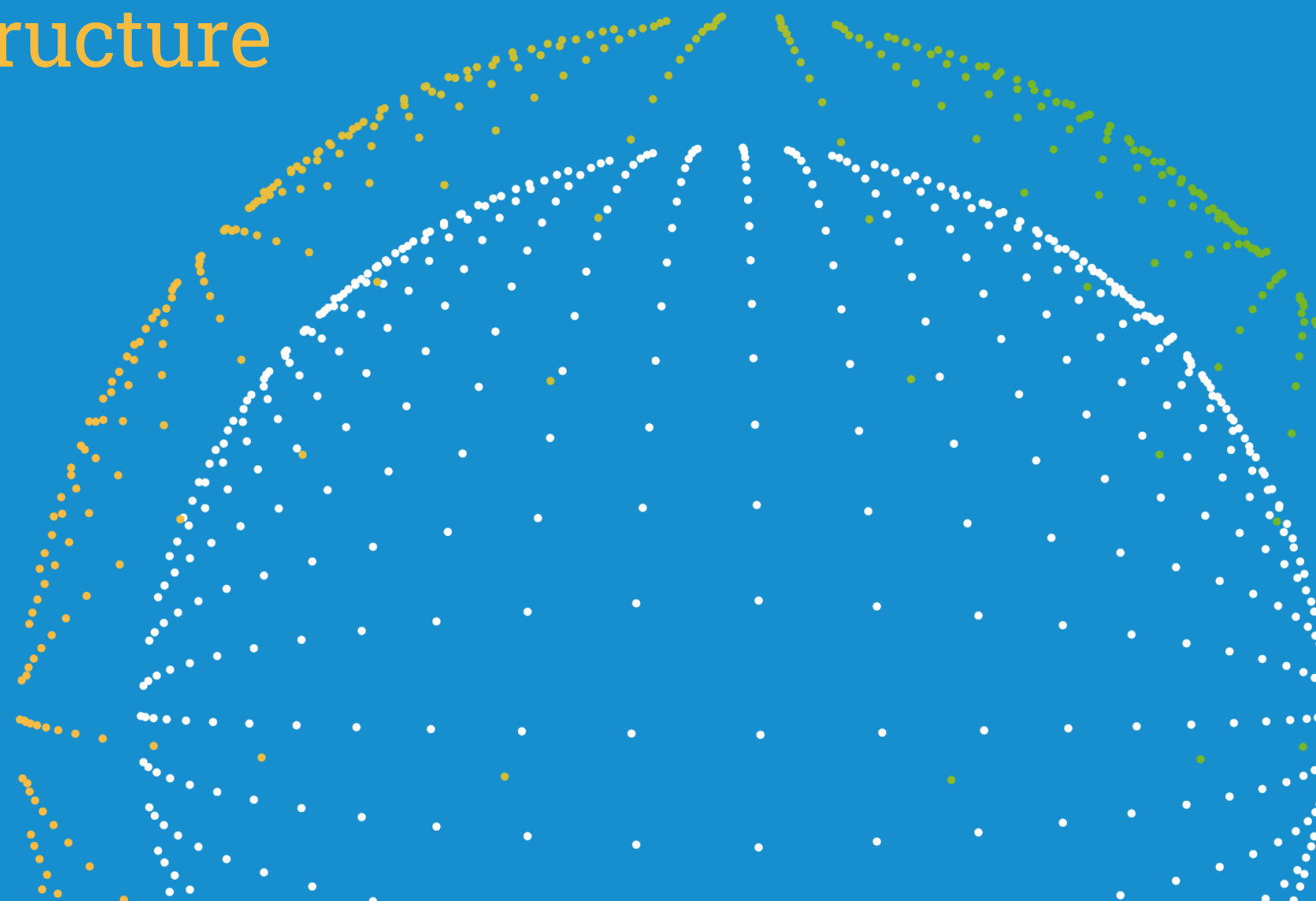
13-03-2024 Data Spaces Symposium, Session: Data, AI and beyond in language, cultural heritage and media
<https://language-data-space.ec.europa.eu>

Data Spaces Symposium

Towards sustainable language data
and services infrastructure

Andrejs Vasiļjevs

Co-Founder and Board Member, TILDE
Member of the Board of Directors, BDVA





Generative AI is powered by large language models (LLMs). Language means culture, and we've got a very diverse set of cultures and languages in Europe. Therefore, one cannot only speak about global generative AI applications. They should also be localized into the local context and culture. And if anything, Europe is quite a unique place in which to do so.

McKinsey
& Company

Source: KcKinsey&Company, "Leveraging generative AI in Europe: The opportunities and challenges"

Search & exchange language resources

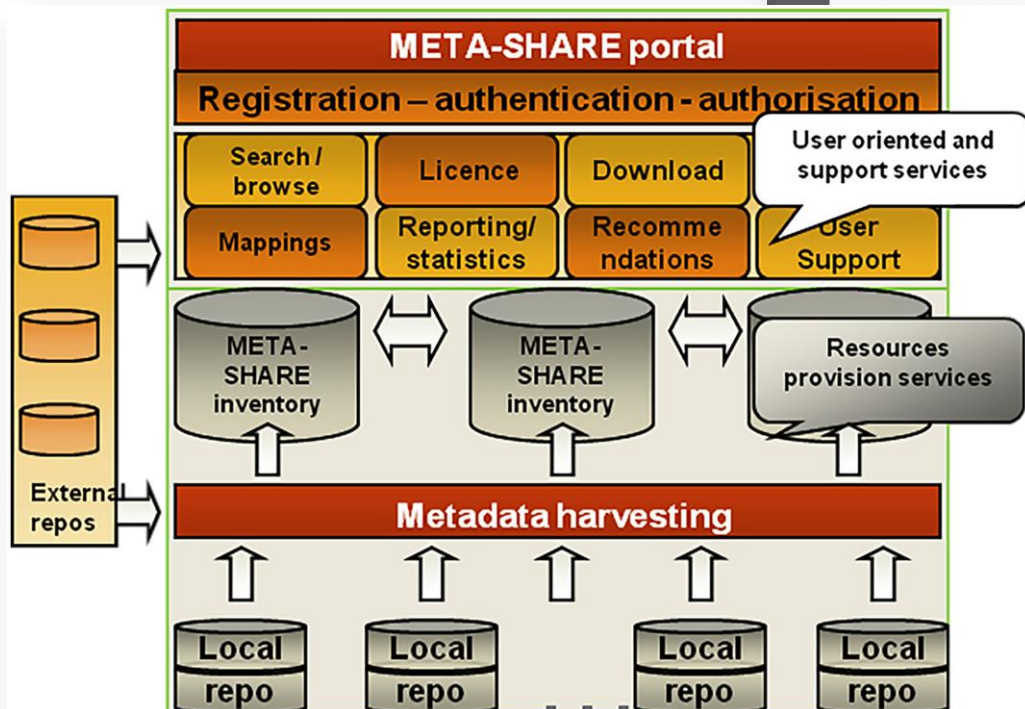
META-SHARE is an open and secure network of repositories for sharing and exchanging language data, tools and related web services

Share your own resources!

[JOIN OUR NETWORK NOW](#)

Already a member? [Log in](#)

Search the META-SHARE inventory



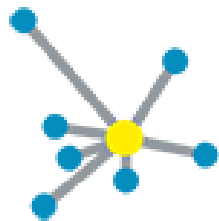
4,481 users

2,887 language resources

32% text corpora

27,630 number of downloads

A Network of Excellence consisting of 60 research centres from 34 countries dedicated to building the technological foundations of a multilingual European information society



CEF, 2014-2023

Identifying,
collecting, and
sharing
language data
from public
administrations

Filter by:

▼ Resource Type

▬ Corpus (69)

▸ Language

▸ Media Type

▸ Licence

▸ Conditions of Use

▸ Linguality Type

▸ Multilinguality Type

▸ Data Format

▸ Domain

▸ Appropriateness For DSI

▸ Funding Project



ELRC-SHARE Repository



-  **Anonymised ParaCrawl release 8 Latvian-English** 0 28
English | Latvian CC0-1.0
-  **Anonymised ParaCrawl release 9 English-Latvian** 0 29
English | Latvian CC0-1.0
-  **CEF Data Marketplace multilingual benchmark for the evaluation of cleaning and clustering tools** 26 74
Czech | English | German | Italian | Latvian



H2020, 2019-2022



One Platform for all European 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

ELG RELEASE 3.0 (MARCH 2024) — GRID WORKING — NO MAINTENANCE SCHEDULED

8073
Corpora

3849
Tools & Services

2823
Conceptual Resources

512
Models & Grammars

1779
Organizations

514
Projects



Tilde MT Machine Translation engine, Latvian - English

Tilde MT LV-EN

Version: 1.0.0 (automatically assigned)

ELG-compatible service (service running on the provider's side)

Keyword

Machine Translation Neural MT Latvian English

Intended application

Machine Translation

Overview Download/Run Try out Code samples

Original text

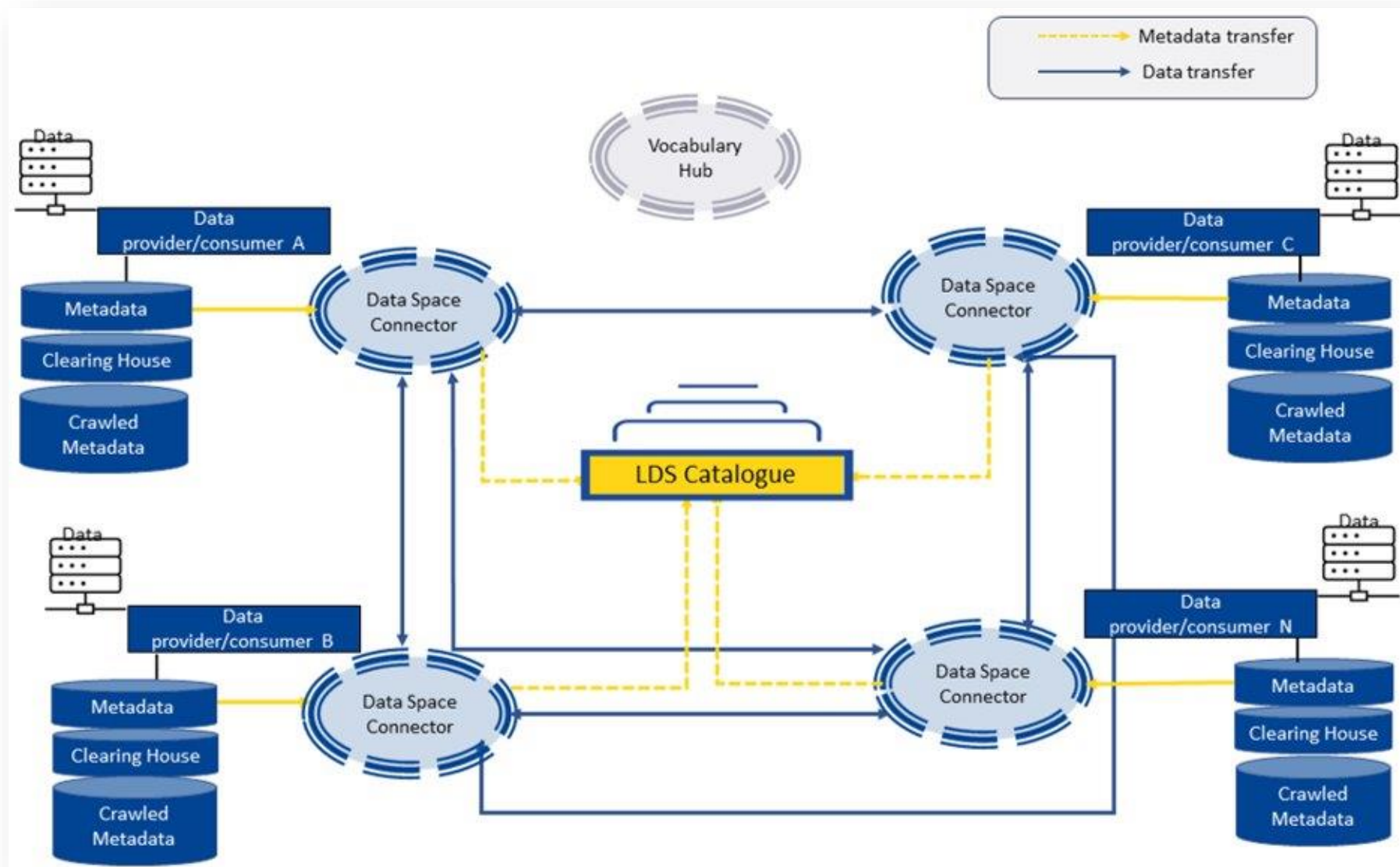
Valodas resursiem ir nozīmīga loma mākslīgā intelekta sistēmu izveidē.

Translated

Score: 0
Language resources play an important role in the development of artificial intelligence systems.

BACK

The platform for European language technologies, resources, and services, fostering collaboration and innovation across the Europe's language technology sector.

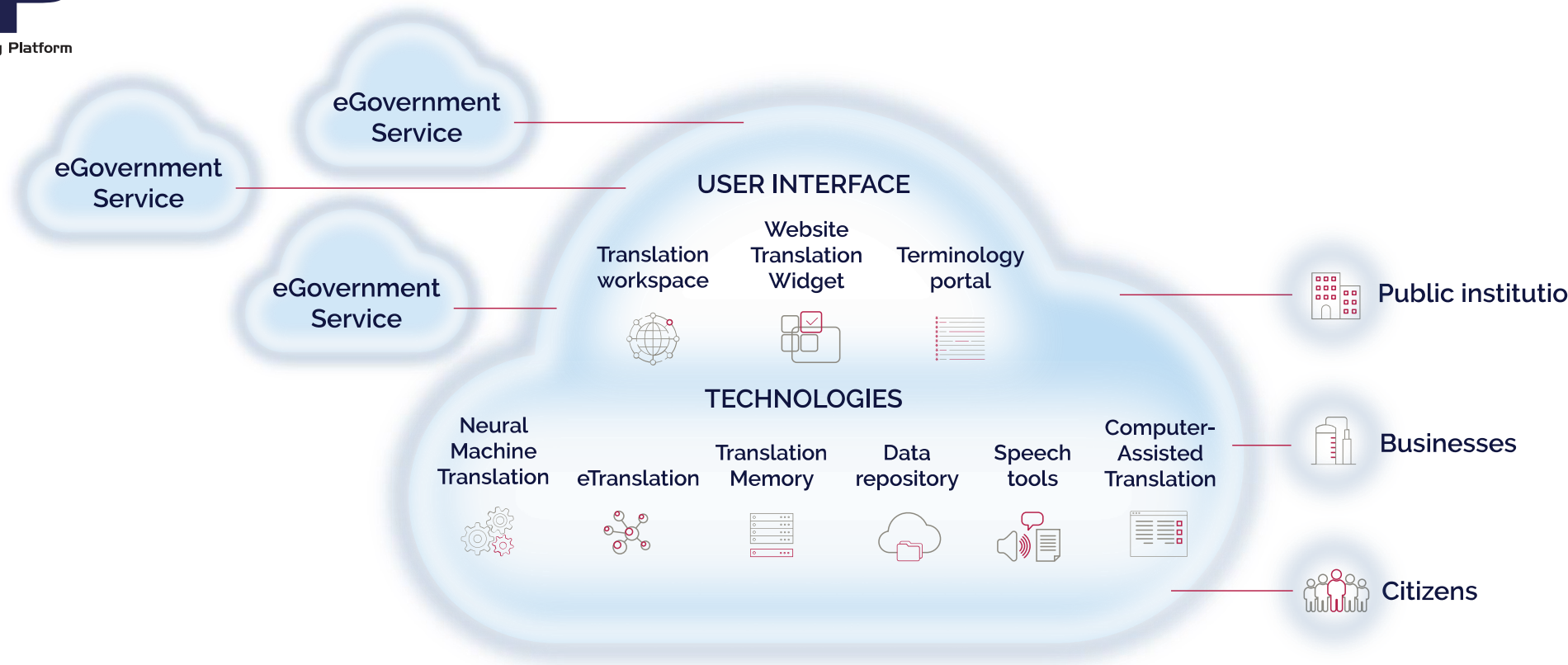


Federated language data sharing infrastructure aimed at supporting the digital language equality in the European Union by facilitating access to, sharing, and use of language data and services.

ALT-EDIC

The European Alliance for Language Technologies

- ALT-EDIC will pool Member States' funding and other resources in a flexible and efficient way, to invest in transformative work on language data and language technologies.
 - European Digital Infrastructure Consortium (EDIC) is a new mechanism for multi-country projects of critical importance, adopted by European Commission as part of AI Innovation Package on Jan 24, 2024
- The ALT-EDIC's mission is to develop a common European infrastructure in Language Technologies, particularly Large Language and other foundational AI models. It seeks to improve European competitiveness, increase European data and other relevant resources and uphold Europe's linguistic diversity and cultural richness:
 - Language Data
 - Existing language tools and language models
 - New Language tools and Large Language Models
 - Evaluation, certification, and normalization
 - Language Ecosystem



Open platform that provides services, tools, and resources to advance the use of language technologies tailored to the nation's linguistic needs.



Latvia



Estonia



Malta



Croatia

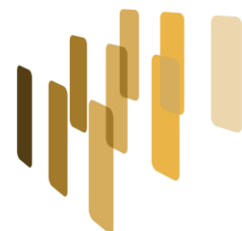


Iceland

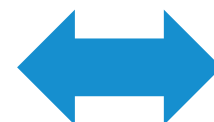


Towards sustainable language data and services infrastructure

META SHARE



EUROPEAN
LANGUAGE
GRID



EUROPEAN
LANGUAGE
DATA SPACE

European Language
Resource Coordination
Connecting Europe Facility
ELRC-SHARE



National and local initiatives



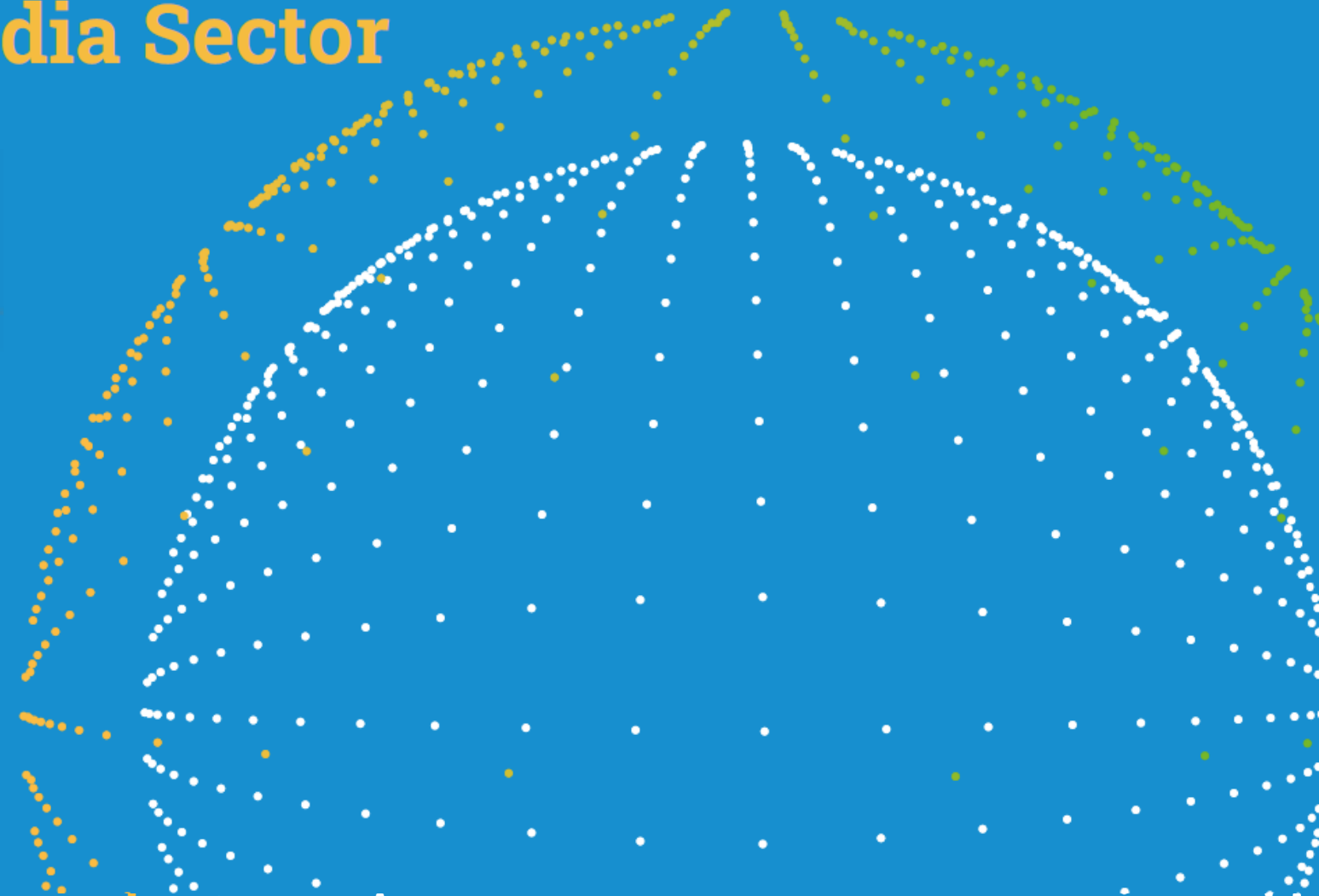
ALT-EDIC

Data Spaces Symposium

Data Sharing for new Business
opportunities in Media Sector



Oscar Rey
Innovalia Association
13/03/2024





www.tems-dataspace.eu



Funded by the Digital Europe Programme of the European Union Under Grant Agreement No 101123423

TEMS EU DEPLOYMENT

Project No: 101123423

Start Date: 1st October 2023

Duration: 36 Months

Partnership: 32 Full Partners, 7 Associated Partners, 4 Affiliated Entities, 13 Countries

Strategic Objective: Set up and deploy a secure and trusted data space to enable Audiovisual and Media organizations to cooperate by sharing and accessing data in a mutually advantageous manner and in full compliance with the data protection legislation.

Total Budget: 16,5 MEUR



Media Data Driven Processes

Broadcasters, Publishers, News Agencies, Producers, Audiovisual national archives

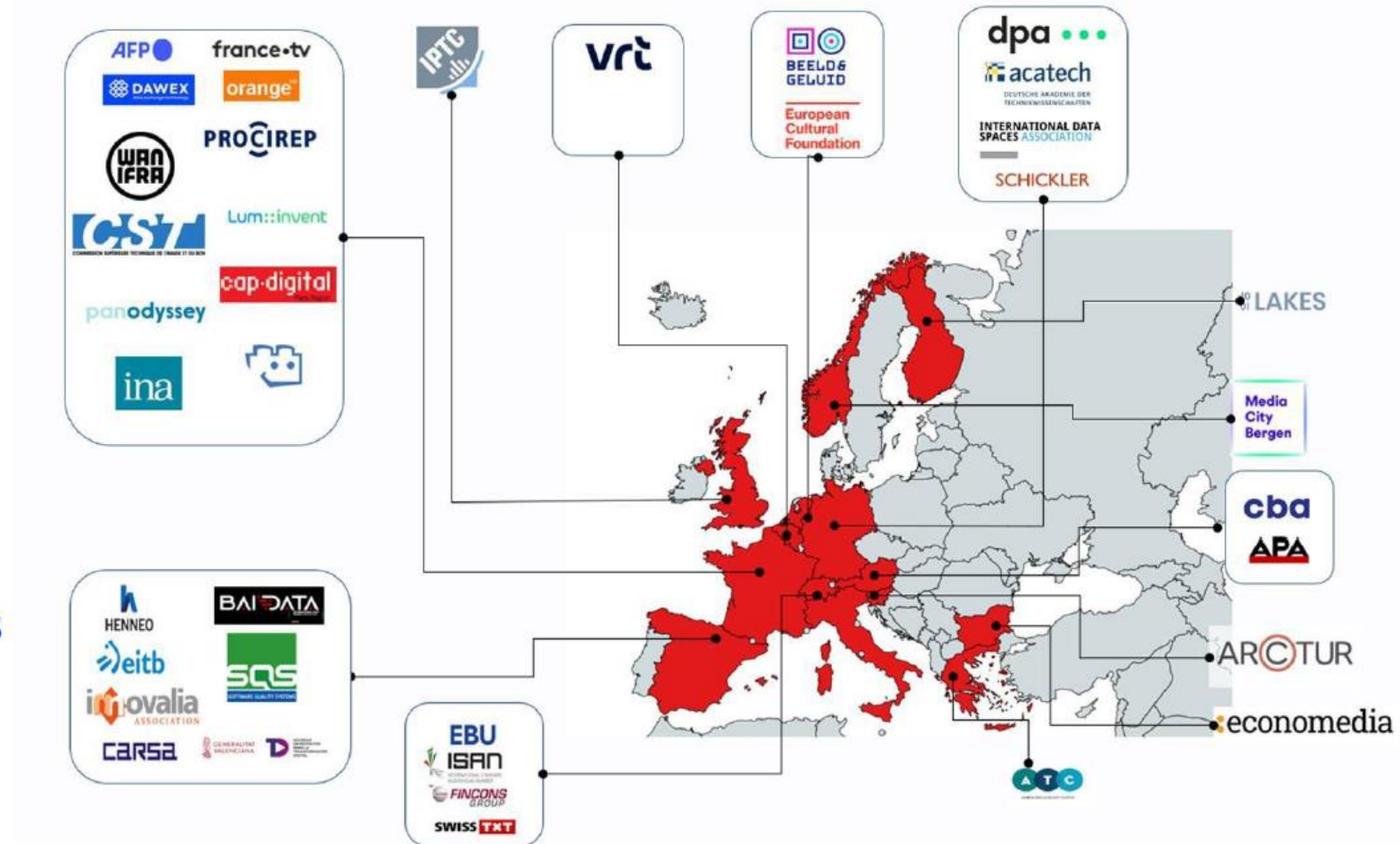


FIGHTING MISINFORMATION

AUDIENCE ANALYSIS

IMPROVING DATA FLOWS IN PRODUCTION CHAINS

SUPPORTING THE ADOPTION OF AI AND VIRTUAL REALITY TECHNOLOGIES

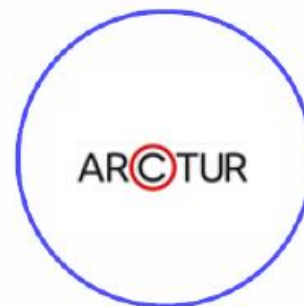


IT providers in data space design and operations

COLLABORATIVE DATA DRIVEN MEDIA VALUE CHAINS TRIALS



NEWS AND FACT-CHECKING, AUDIENCE DATA, PERSONALIZATION & REVENUE
STREAMS, COLLABORATION IN PRODUCTION CHAIN AND DRM, INNOVATION &
NEW MEDIA FORMATS



DATA SHARING: fact-checks, news content, media usage data, 3D data for virtual production, metadata for audiovisual works, etc.

BENEFITS FOR THE MEDIA SECTOR: Credibility, Personalize content, Innovate in audiovisual production, Interoperability, new revenue streams, Collaboration and efficiency and co-creation in 3D environments.

USE OF AI:

- Reliable source for fact-checking in AI generative tools.
- AI-driven personalization service for enhancing the user experience (News).
- Categorizing and analyzing textual information associated with audiovisual productions.
- Transcripts and translations.
- IPR Management.
- Discovery and utilization of 3D data.
-



DATA SHARING: fact-checks, news content, media usage data, 3D data for virtual production, metadata for audiovisual works, etc.

BENEFITS FOR THE MEDIA SECTOR: Credibility, Personalize content, Innovate in audiovisual production, Interoperability, new revenue streams, Collaboration and efficiency and co-creation in 3D environments.

USE OF AI:

- Reliable source for fact-checking in AI generative tools.
- AI-driven personalization service for enhancing the user experience (News).
- Categorizing and analyzing textual information associated with audiovisual productions.
- Transcripts and translations.
- IPR Management.
- Discovery and utilization of 3D data.
-



Data Spaces Symposium

TEMS:Use cases description.

Sylvain Le Bon
Startin' Blox

An abstract graphic of a globe or sphere composed of numerous small dots. The dots are primarily white, with some yellow and orange dots scattered throughout. The dots are arranged in a grid-like pattern that follows the curvature of the sphere, creating a sense of depth and perspective. The background is a solid blue color.

PILOT 1: B2B exchange platform for fact-checking and news content



Pilot description: Sharing news content produced by certified fact-checkers, news media and other media organizations including other B2B stakeholders.

Objective: Increase visibility/exposure and/or knowledge transfer, and/or monetization through republication or AI use.

AI: Enhance news verification tools by utilizing TEMS as a reliable source for fact-checking and news content in AI generative tools, such as ChatGPT.

Media and Techs involved: AFP, APA, DPA, EDMO (European Digital Media observatory), VRT, NISV.

PILOT 2: Marketplace for content and services



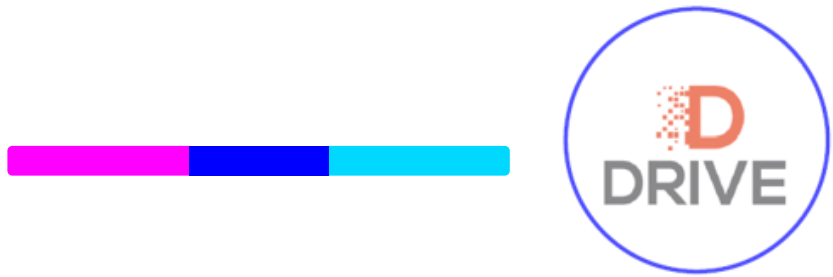
Pilot description: Distributed marketplace integrated in TEMS Data Space for content syndication and media monitoring services.

Objective: Gain visibility and attracting larger media outlets for syndication, leading to increased revenue opportunities.

AI : Enhance the efficiency of searching and acquiring datasets within EMDS, providing users with more relevant and personalized content recommendations.

Media and Techs involved: APA and (indirectly) Austrian publishers

PILOT 3: Next level DRIVE – The Digital Revenue Initiative



Pilot description: TEMS as a centralized environment for the collection, sharing, and analysis of data related to media trends and personalization services.

Objective: Enhance media companies' capabilities in understanding media trends and providing personalized digital news products

AI : AI-driven personalization service for enhancing the user experience by delivering more tailored and relevant news content to individual end-users.

Media and Techs involved: Deutsche Presse-Agentur and Schickler, other newspaper and media publishers

PILOT 5: Collaboration in Production Chain



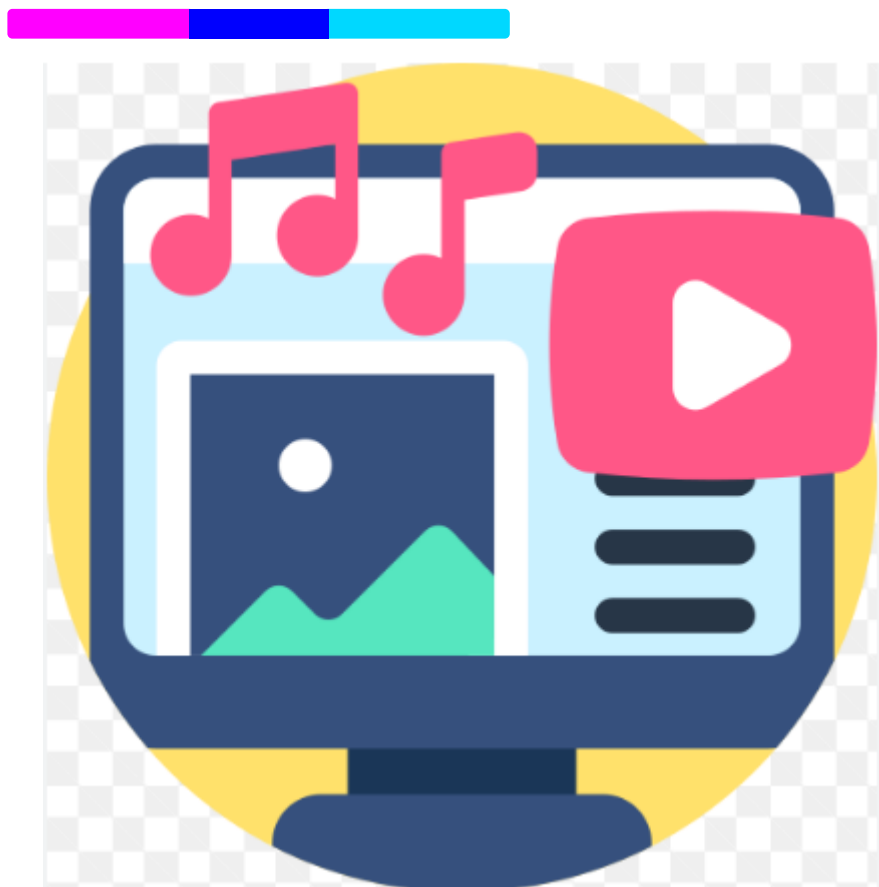
Pilot description: TEMS as an decentralized data exchange framework as a service for the audiovisual production ecosystem

Objective: Enhancing interoperability and tracking of metadata associated with audiovisual works.

AI : Ensure the preparation of data sets of good quality that could be exploited for algorithms' training, thanks to the better data tracking. AI based tools will contribute to enrich descriptive metadata linked to an AV work, enabling a better "findability".

Media and Techs involved: Cap Digital, CST, France Télévisions, France Télévisions Studio, INA, PROCIREP, SPTD // ISAN-IA, Lum::Invent, Startin'Blox, Perfect Memory.

PILOT 6: B2B exchange platform for content syndication



Pilot description: Sharing media content(text, audio, and video) produced by community and independent media organizations including other B2B stakeholders.

Objective: Increasing visibility/exposure and/or knowledge transfer, and/or monetization through republication.

AI :Facilitate the cross-language search functionality, allowing personalized recommendations, enabling the automatic generation of transcripts and translations, etc.

Media and Techs involved: Display Europe, Krytyka Polityczna, OktoTV, Cultural Broadcasting Archive (cba.media)

PILOT 7: Syndication platform for written works



Pilot description: Creating interoperable connectors, standardizing data exchange, and addressing challenges in the digital publishing sector (Writings)

Objective: Enhance the competitiveness of media and cultural publishers in the digital sector.

AI: Enhancing intellectual property value & promoting transparency and ethics

Media and Techs involved: Publishers and Press Agencies in the media and cultural sectors

PILOT 8: Market and co-creation place for 3D environments for virtual production



Pilot description: European market and co-creation platform within TEMS, facilitating the discovery, collaboration, and storage of diverse 3D data types.

Objective: Support the development and re-use of innovative media formats in virtual production

AI : AI to enhance the discovery and utilization of 3D data, optimizing the creation and co-creation processes for innovative media formats in virtual production

Media and Techs involved: VRT, Arctur, Sound Holding.

Data Spaces Symposium

Europeana and the Data Space
for Cultural Heritage

Valentine Charles



WELCOME TO THE

Common European data space for cultural heritage

THE MERMAID (DANISH: HAVFRUE) IS A BRONZE SCULPTURE DESIGNED BY THE DANISH SCULPTOR ANNE CARL-NIELSEN, DEPICTING A MERMAID.

PLACE: SMK - STATENS MUSEUM FOR KUNST

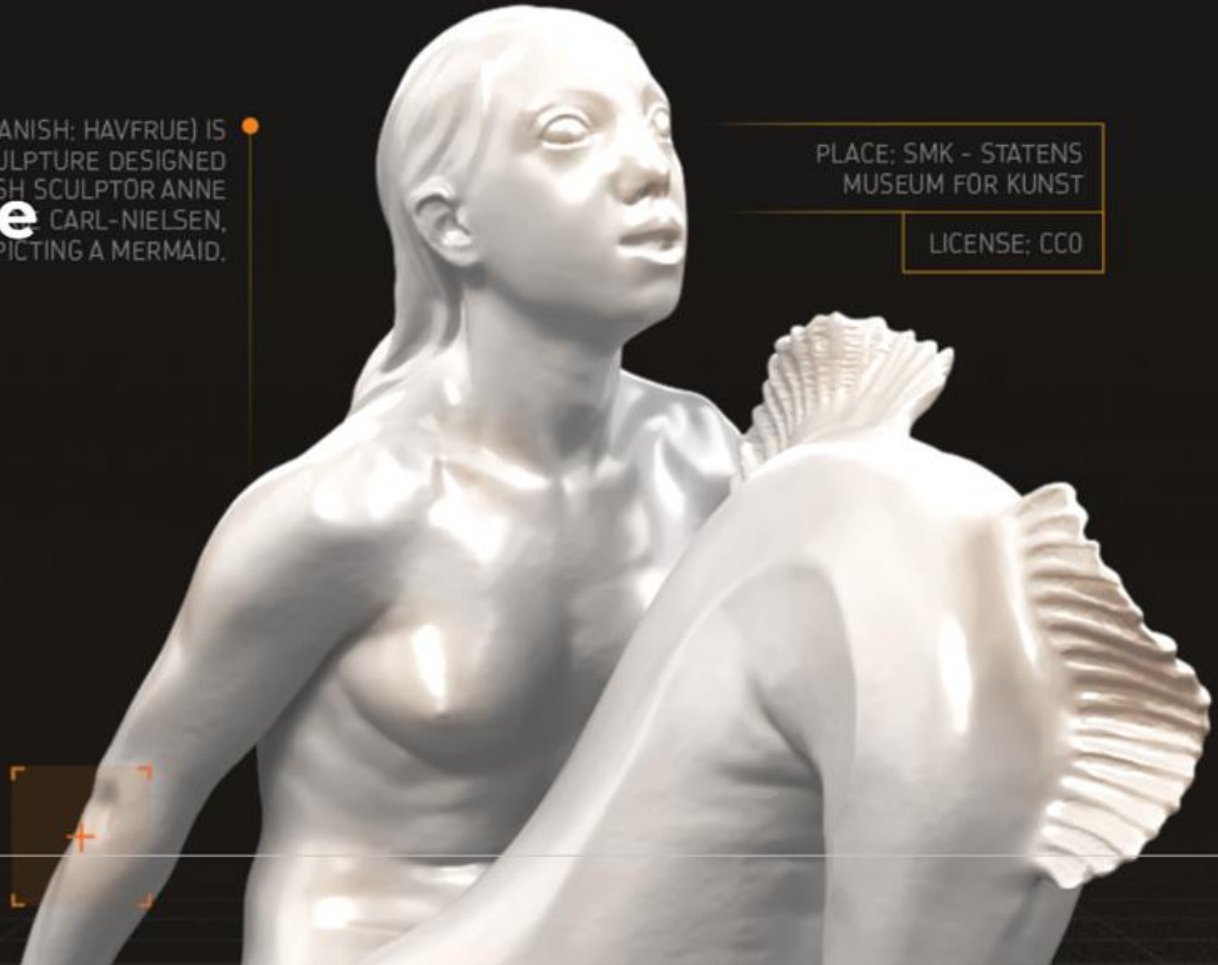
LICENSE: CC0

Access cultural heritage data from across Europe

BROUGHT TO YOU BY



1921



<https://www.europeana.eu/en/dataspace-culturalheritage>

DATA SPACE IN NUMBERS

Millions of items and thousands of collaborations

57,000,000 +

ITEMS

4,500 +

NETWORK MEMBERS

2,600 +

PROVIDING INSTITUTIONS

+10%

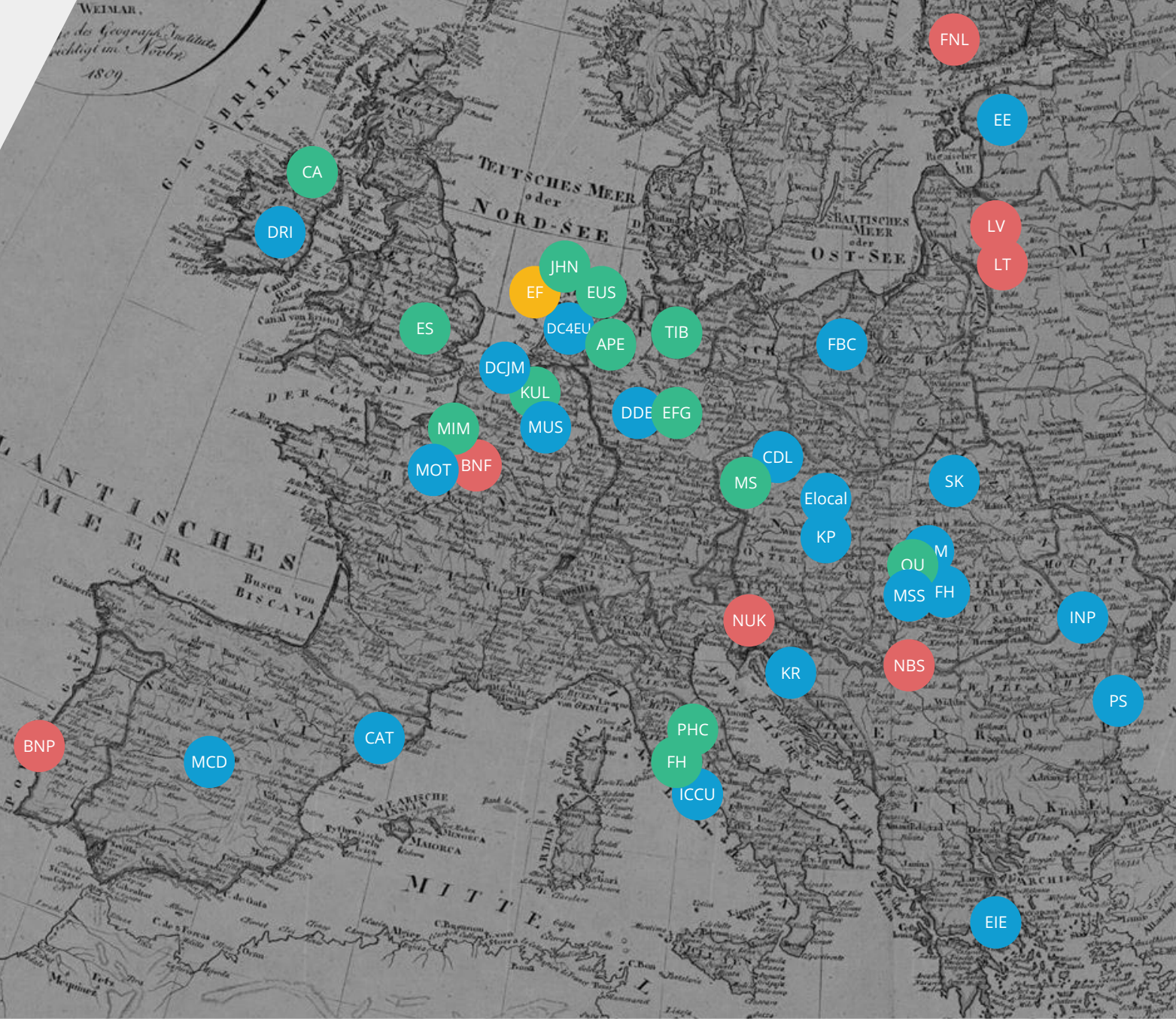
INCREASE IN HIGH-QUALITY
DATA PER YEAR

13,000,000

AVERAGE NUMBER OF
MONTHLY API REQUESTS

42

Accredited aggregators in the European Aggregators' Forum



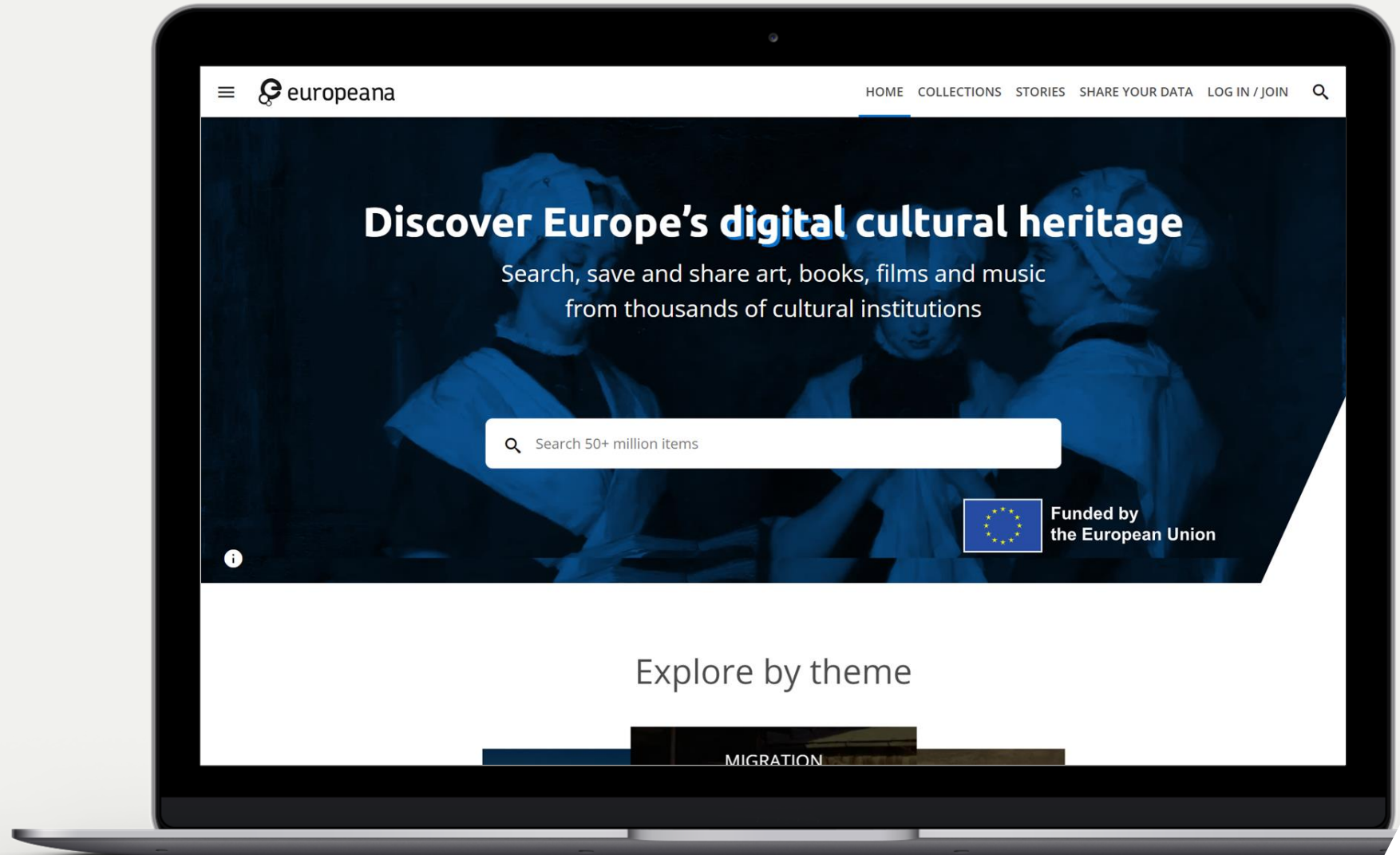
= National Aggregator

= Domain Aggregator

= National Aggregator & National Library



EUROPEANA.EU AND APIs SUITE



AI TO SUPPORT THE USE AND ACCESS TO CULTURAL HERITAGE

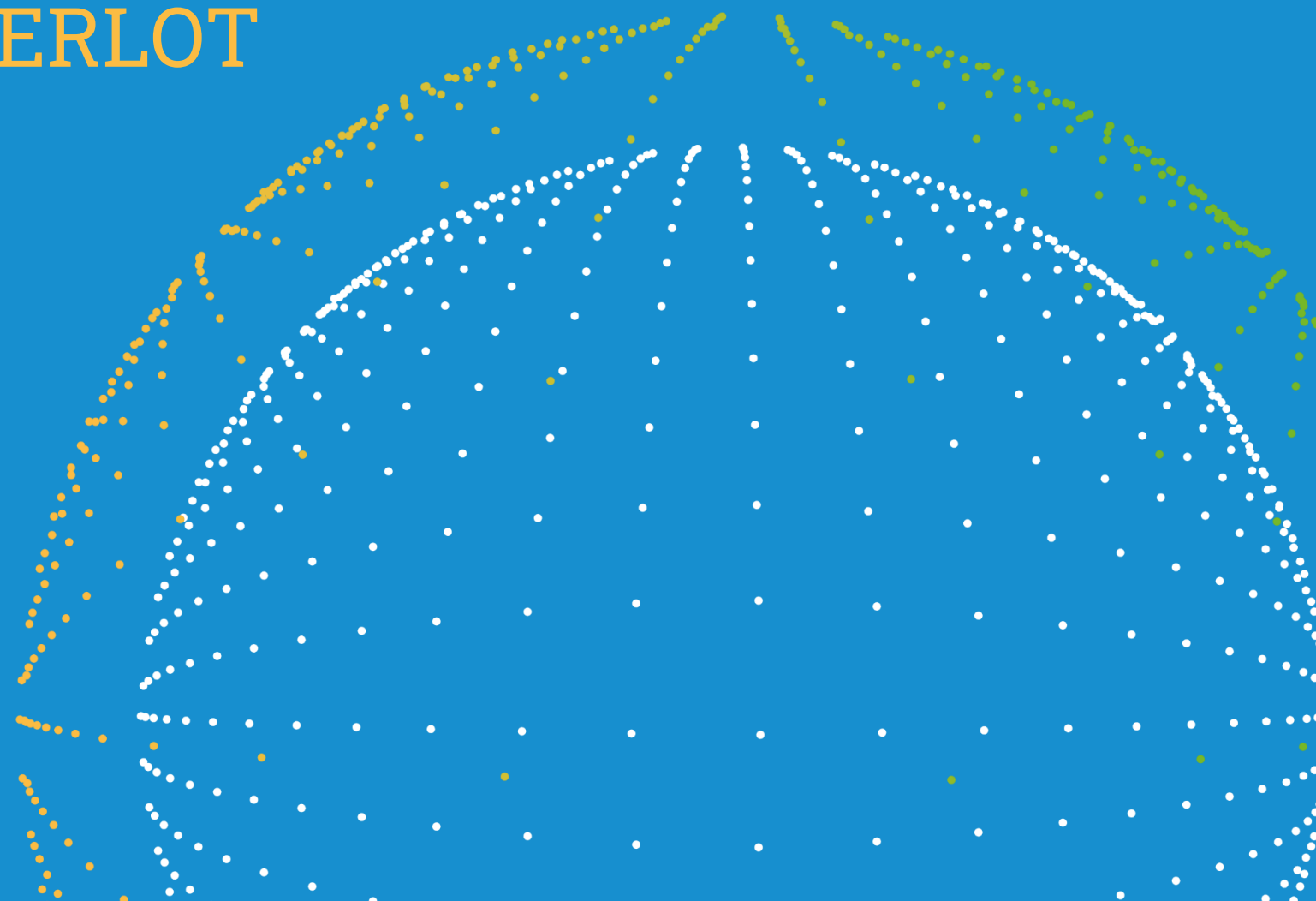
- AI to support the enhancement of data quality
- AI to support the engagement of users on Europeana.eu
- Build capacity around AI (AI4Culture platform to be launched this year)



Data Spaces Symposium

German Gaia-X education
dataspace project MERLOT

Dr. Sabine Zander, imc

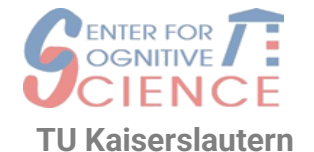


German Gaia-X education dataspace project MERLOT



MarkEtplace foR LifelOng educaTional dataspaces and smart service provisioning (MERLOT)

- **Funded by:** German Federal Ministry for Economic Affairs and Climate Action
- **Coordinator:** imc AG
- **Consortium partners:** 11
- **Runtime:** January 2022 – December 2024



The need for upskilling and reskilling



The 2023 World Economic Forum's Future of Jobs Report predicts that **61 percent of employees will have to be retrained by 2027.**

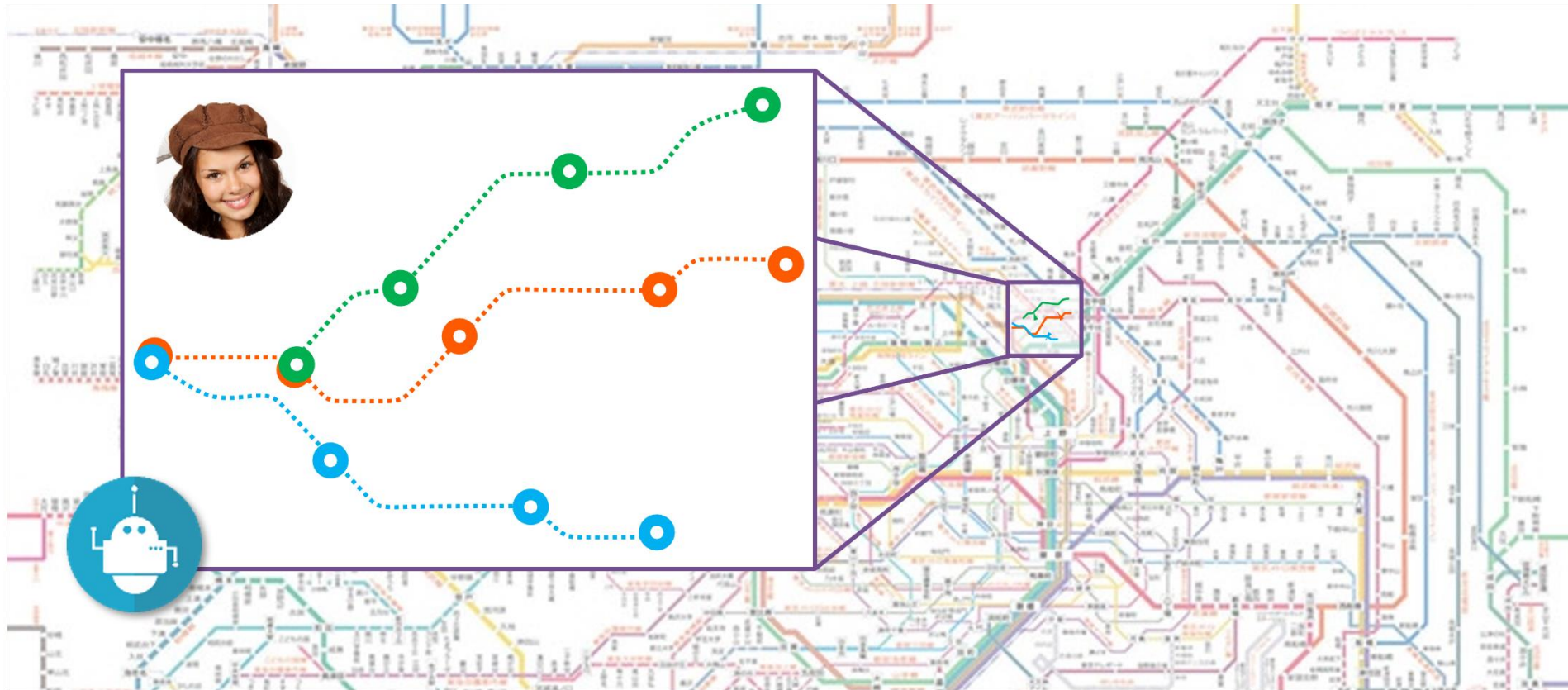


Workers want to reskill. 77% are ready to learn new skills or completely retrain (PwC's Global Workforce Hopes and Fears Survey 2023).



Gartner research shows that HR technology leaders have identified **skills management as one of the most important HR technologies** (2023 survey).

Lifelong learning needs data (AI)



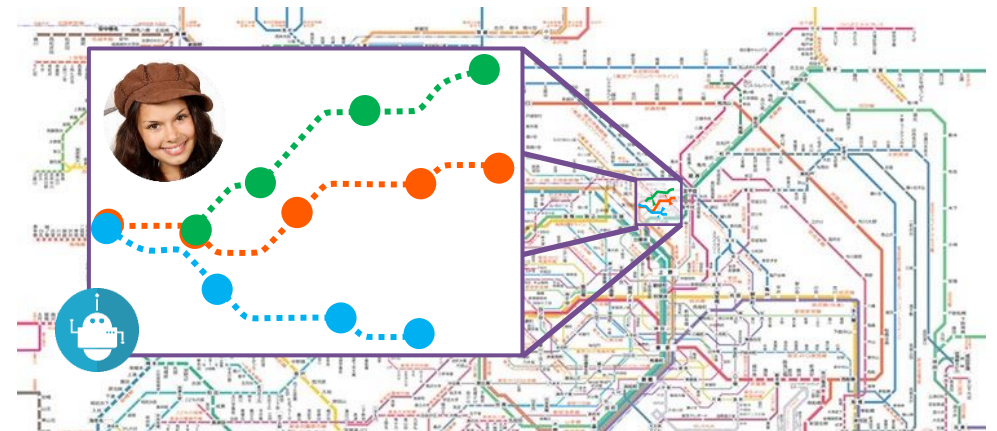


From *data silos* to *data spaces*

Personal education related data is **highly vulnerable**, therefore locked in *data silos* of institutions



AI education revolution through personal adaptive tools and assistance depends on **access to data** in **secure data spaces**



MERLOT Project Goals



1. From data protection to data sovereignty for individuals
2. Guarantee data sovereign data sharing for organizations
3. Interoperability of data and services
4. Sustainable dynamic development of the education data economy and enabling data-driven business models
5. Development of innovative education services



Data Spaces Symposium

Unite. Innovate. Adopt.

Data, AI and beyond in language, cultural heritage and media sectors

13 March 2024 | 15:45 - 16:55



Andrejs Vasiljevs
Tilde



Oscar Rey
Innovalia Association



Sylvain Le Bon
Startin'blox



Sabine Zander
imc



Valentine Charles
Europeana



Georg Rehm
DFKI



Daniel Alonso
BDVA



Open call for the deployment of the common European data space for skills

DIGITAL-2024-CLOUD-DATA-06-SKILLS-Data Space for Skills (deployment)

The awarded proposal will **integrate, test and deploy the data space for skills**, allowing participants to **make data available and accessible**, as well as **sharing** it, in a controlled, simple and secure way.

Opening date: 29 February 2024

Deadline date: **29 May 2024, 17:00 (CEST)**

- Simple Grant – 50% funding rate
- Budget: 3 Million EUR
- Submit questions in the F&T portal: https://european-union.europa.eu/contact-eu/write-us_en



Open call for the deployment of the common European data space for skills

- **Expected outcomes:** deployment of a Data Space for Skills, including the technical infrastructure, a governance mechanism, continuous maintenance, usage monitoring, helpdesk, sustainability beyond the end of the project, fostering engagement, and three use cases.
- **Targeted stakeholders:** education institutions, HR organisations, employment agencies, public employment services, guidance providers, IT developers, trade and industry associations, alliances and social partners, other private and public actors

Data Spaces Symposium

Unite. Innovate. Adopt. 

Thank you!



Funded by
the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

