EUROPEAN
LANGUAGE
DATA SPACE

Common European Language Data Space

Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiļjevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitris Galanis, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdiņš, Elena Leitner, Valérie Mapelli, Hélène Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault and Leon Voukoutis

22/23/24-05-2024 LREC-COLING 2024
https://language-data-space.ec.europa.eu

Context: Large Language Models (LLMs)

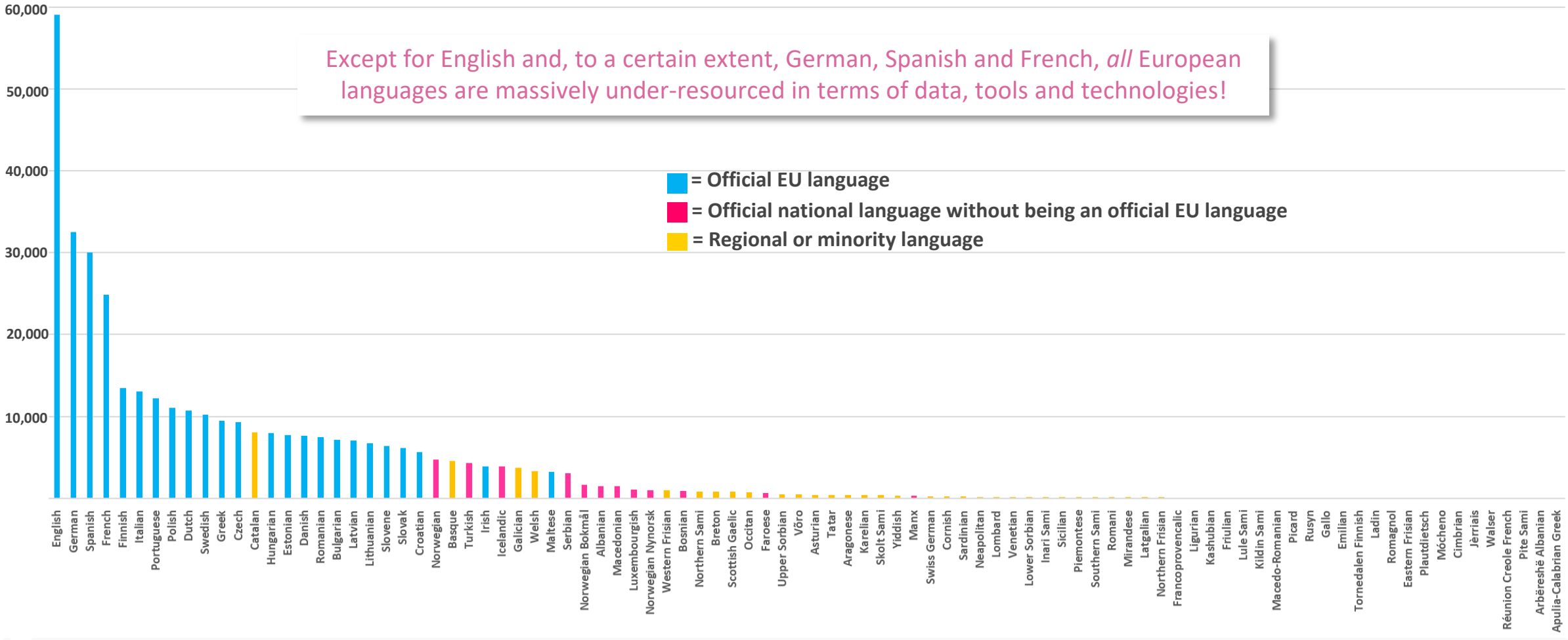
- Large language models are the most disruptive breakthrough in AI in recent history (BERT, GPT-3, ChatGPT, GPT-4 etc.)
- LLMs are trained on vast amounts of training data (language data)
- LLMs use dozens, some even hundreds of terabytes (trillions of tokens) of language and also image, video, audio etc. training data
- The global LT/NLP/AI market is getting bigger and bigger – current assessments evaluate the market size at 500B$ in the next years
- Europe's languages are vastly under-resourced, except English
- A concerted effort for the collection of enormous amounts of language data for all European languages is very much needed

European Initiatives

- There are various European initiatives for the development of LLMs
  - Large research projects in almost every country, e.g., Spain, Denmark, Italy, Germany etc.
  - Companies in many countries, e.g., Finland (Silo.ai), France (Mistral), Germany (Aleph Alpha)
  - EU and nationally funded projects, e.g., HPLT, TrustLLM; new EU calls with substantial budgets
  - New pan-European initiative: ALT-EDIC
- Challenges:
  - Availability of data for European languages
  - HPC facilities
  - Speed of the big tech players in the US and Asia vs. speed of Europe

Digital Language Equality Metric: Technological Scores

EUROPEAN LANGUAGE EQUALITY



EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy
- Data Spaces will help to establish a data economy in Europe
- Various data economy and data infrastructure initiatives in Europe with slightly different goals and individual positioning but conceptual, technical, legal and operational overlap:
  - Data Spaces Business Alliance (DSBA): Gaia-X, IDSA, FIWARE, BDVA
  - EU: DSSC (incl. DSBA), Simpl, approx. 20 data spaces
- The Common European Language Data Space is one of the currently 14 official EU data space projects with a strong focus on industry

Common European Language Data Space

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

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)

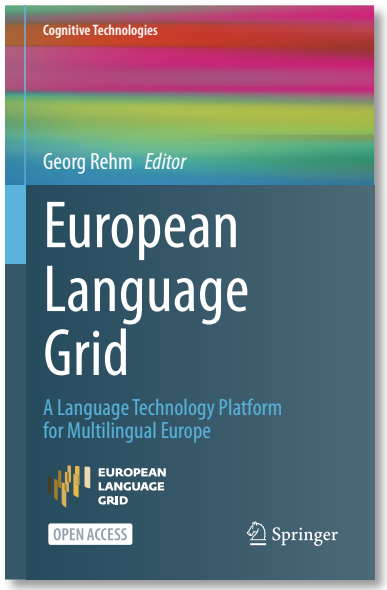
META-NET

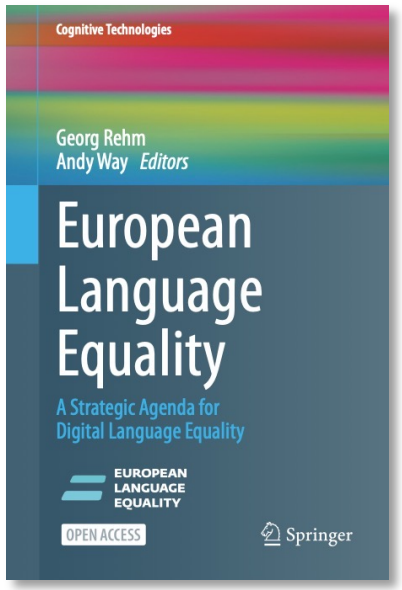
European Language Resource Coordination
Connecting Europe Facility

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY

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



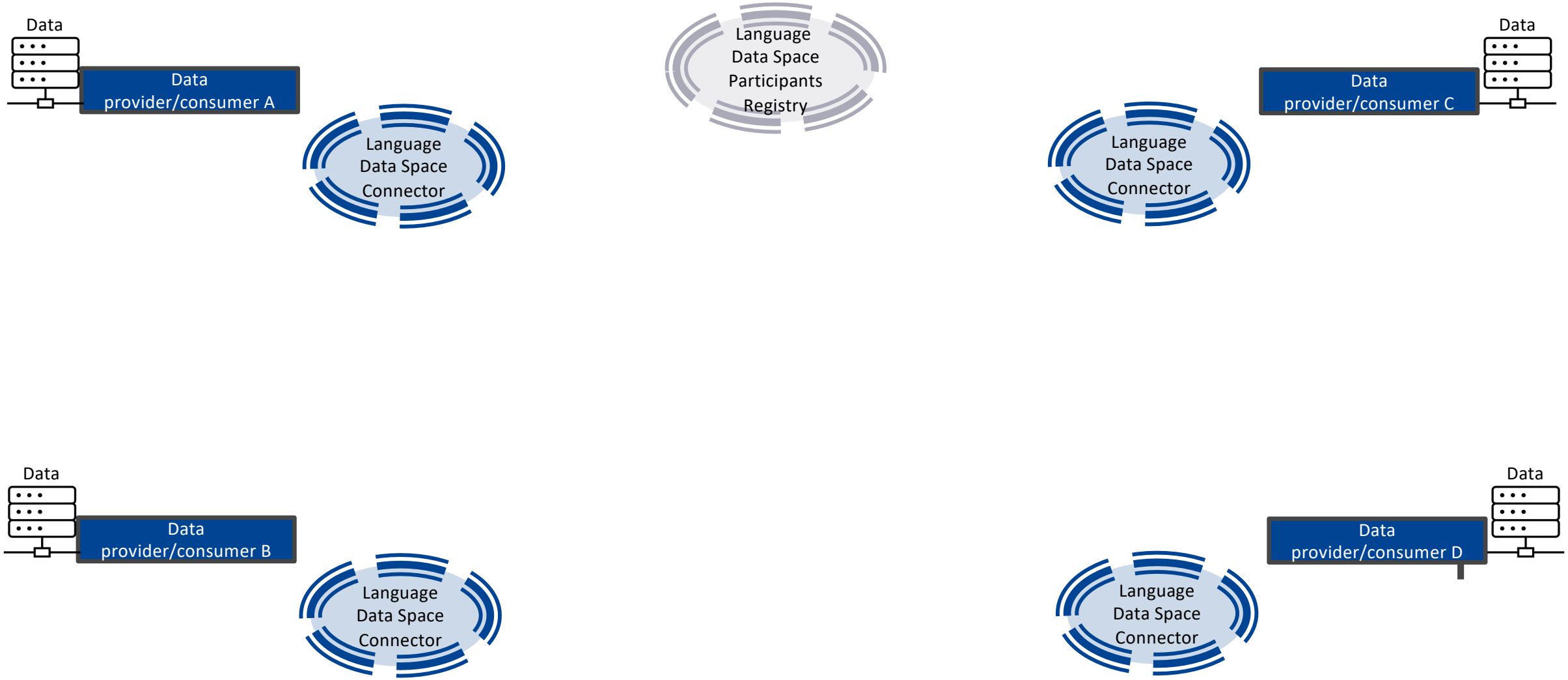


Classes of Data

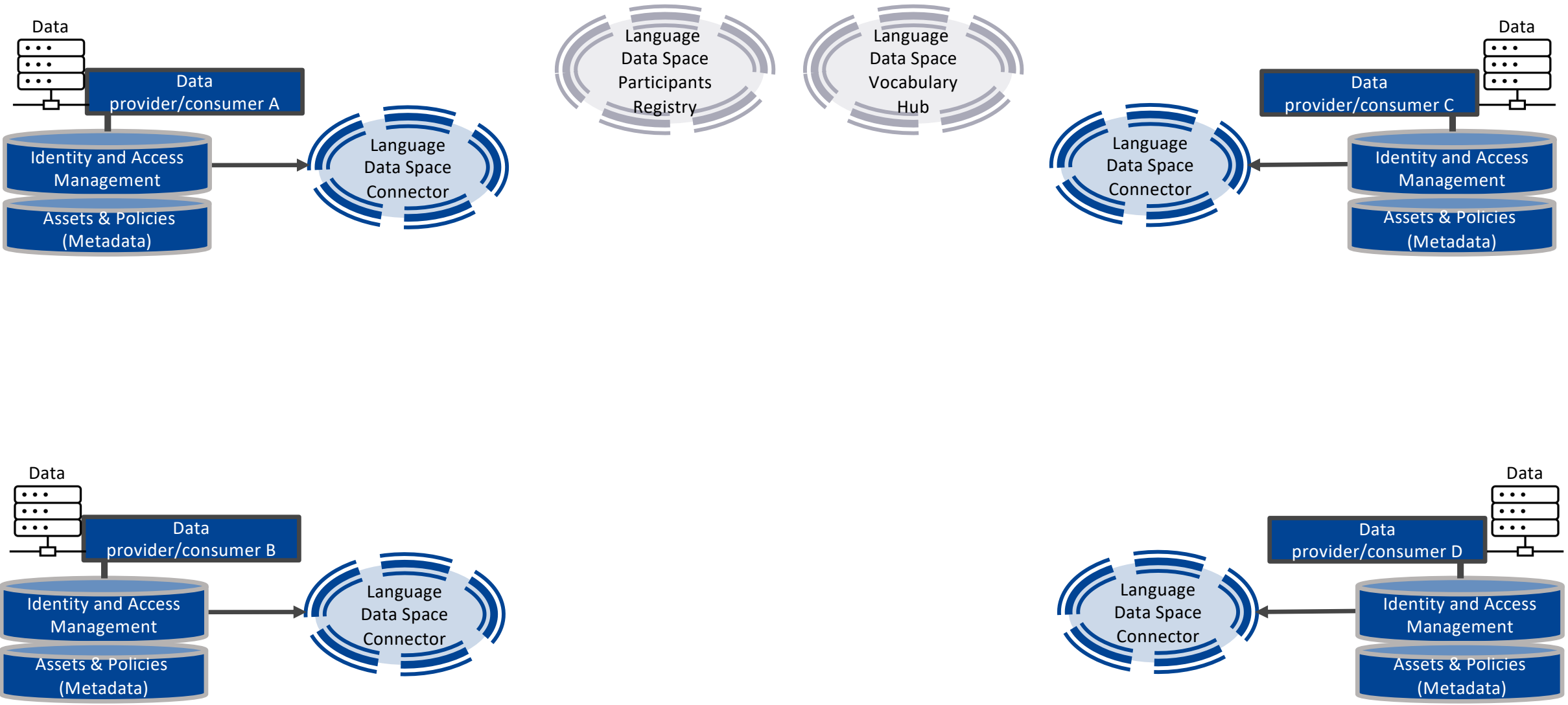
Class of Data	Typical Size	Providers	Integration into LDS	Relevance for LLMs
Regular Corpora and Language Resources	Small (MB, GB)	Primarily NLP/LT research: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Can be easily integrated by connecting the repositories to LDS	Usually very high quality data and thus 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

Alliance for Language Technologies EDIC (ALT-EDIC)

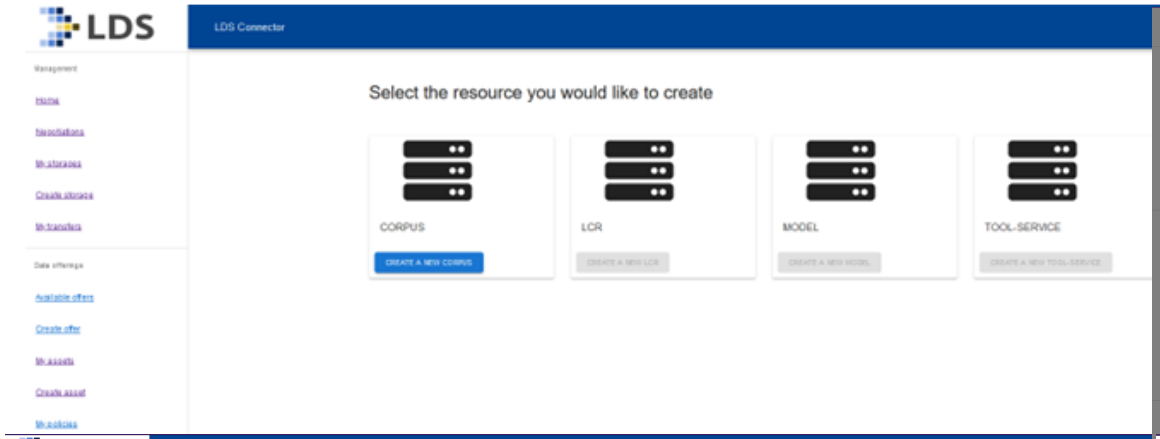
- European Digital Infrastructure Consortium (EDIC): a new legal entity type in the EU
- The first couple of EDICs are currently under development including the ALT-EDIC
- Coordinated by the French Ministry of Culture
- Close collaboration between: ALT-EDIC Working Group, EC, LDS
- ALT-EDIC action plan will concentrate on:
  - 1. Data;
  - 2. Existing language models;
  - 3. New language models;
  - 4. Evaluation, certification, normalization;
  - 5. Ecosystem;
  - 6. EDIC implementation
- We expect many synergies between LDS, ALT-EDIC, DSSC, Simpl, other data spaces and other projects!

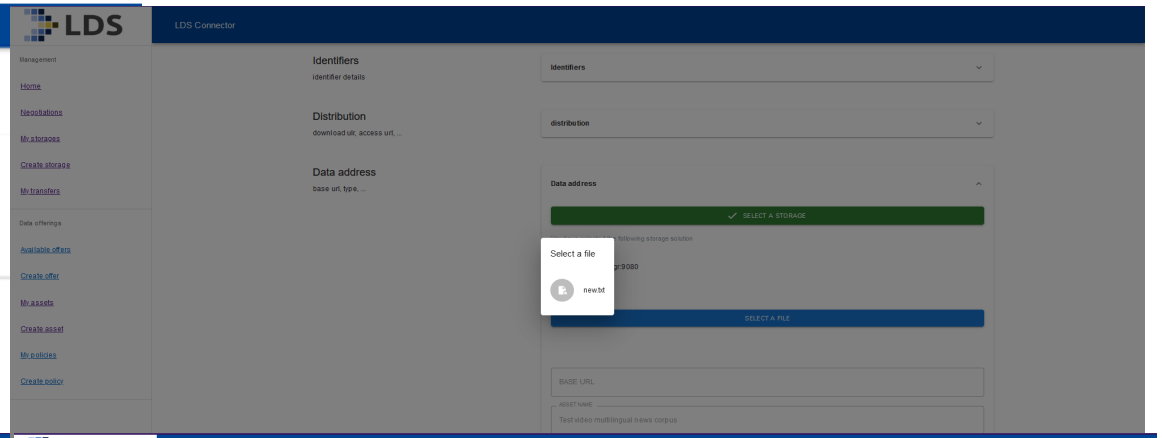

Data
Data provider/consumer A
Language Data Space Connector
Language Data Space Participants Registry
Language Data Space Connector
Data
Data provider/consumer C
Data
Data provider/consumer B
Language Data Space Connector
Language Data Space Connector
Data
Data provider/consumer D

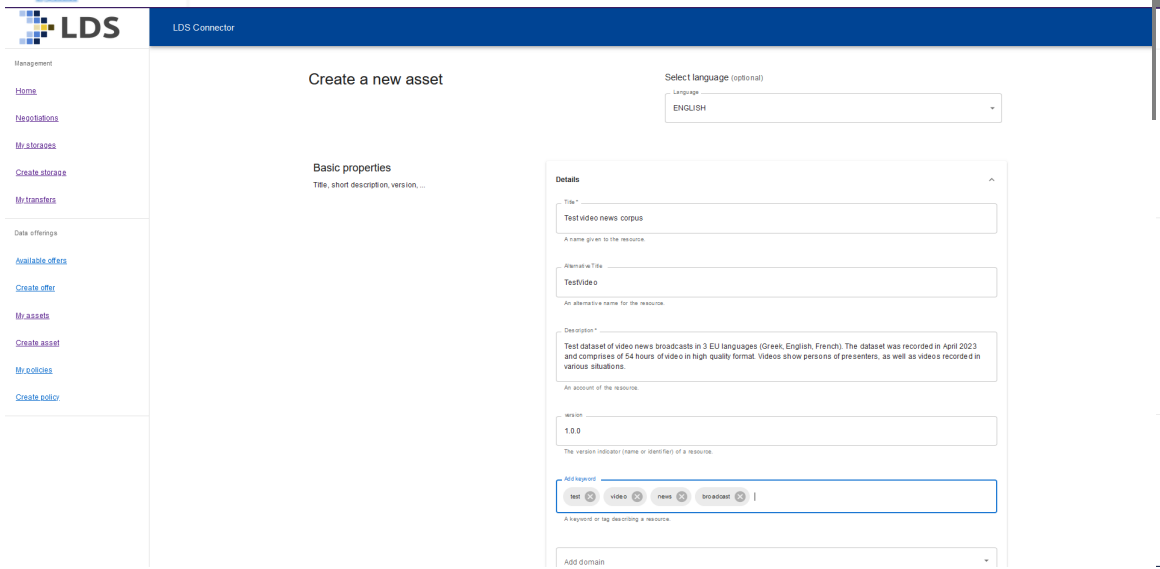

Data
Data provider/consumer A
Identity and Access Management
Language Data Space Connector
Language Data Space Participants Registry
Data
Data provider/consumer C
Identity and Access Management
Language Data Space Connector
Data
Data provider/consumer B
Identity and Access Management
Language Data Space Connector
Data
Data provider/consumer D
Identity and Access Management
Language Data Space Connector

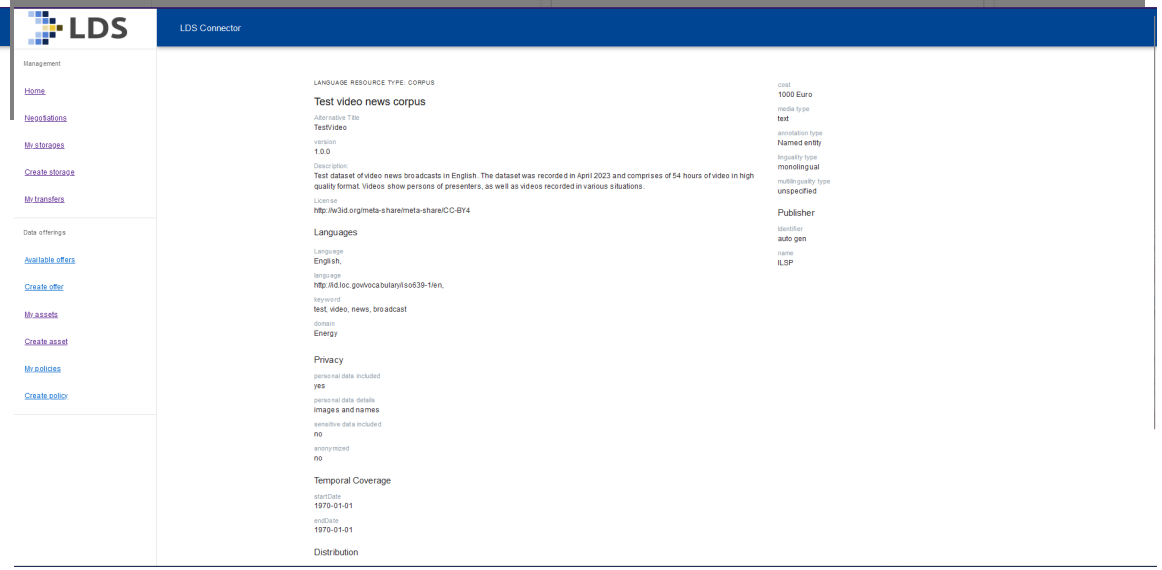

Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Language Data Space Connector
Language Data Space Participants Registry
Language Data Space Vocabulary Hub
Language Data Space Connector
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Language Data Space Connector
Language Data Space Connector
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Data

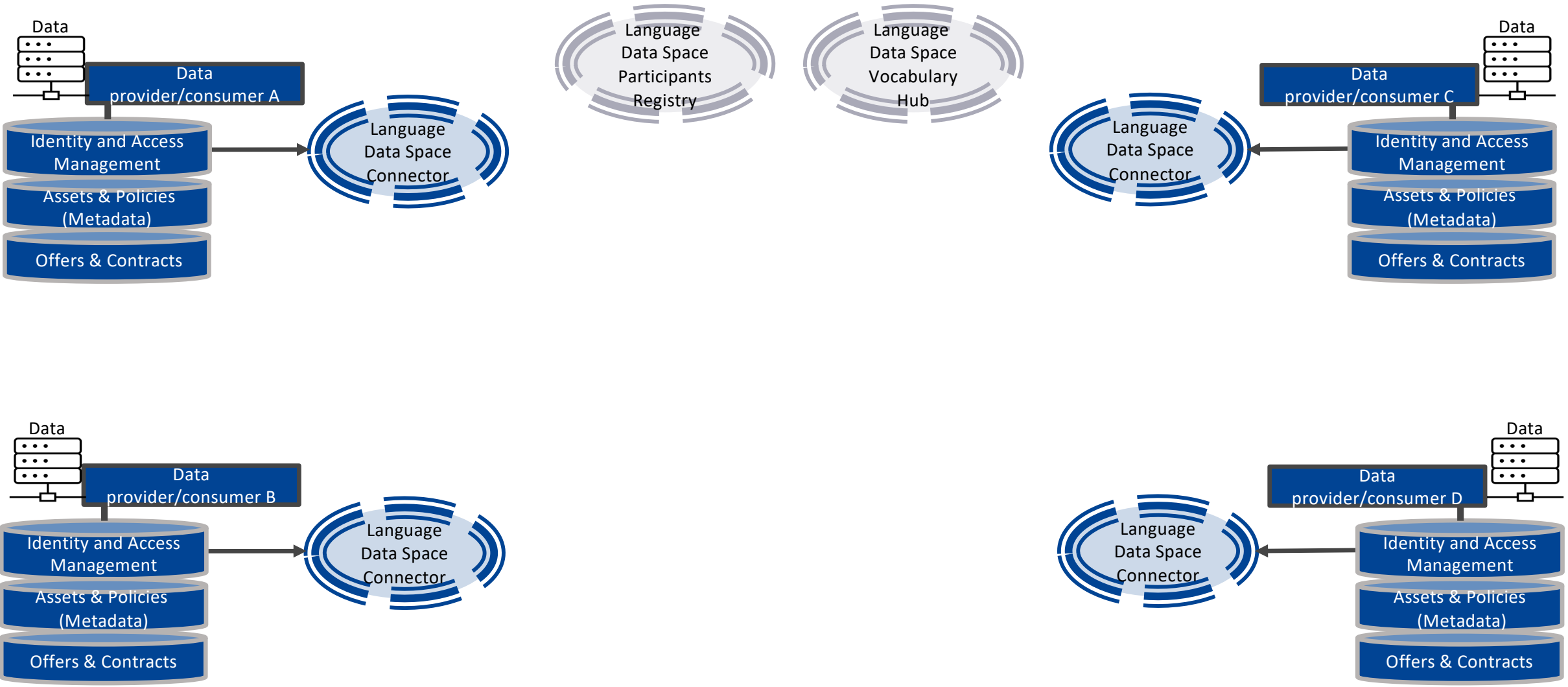
Creating Assets



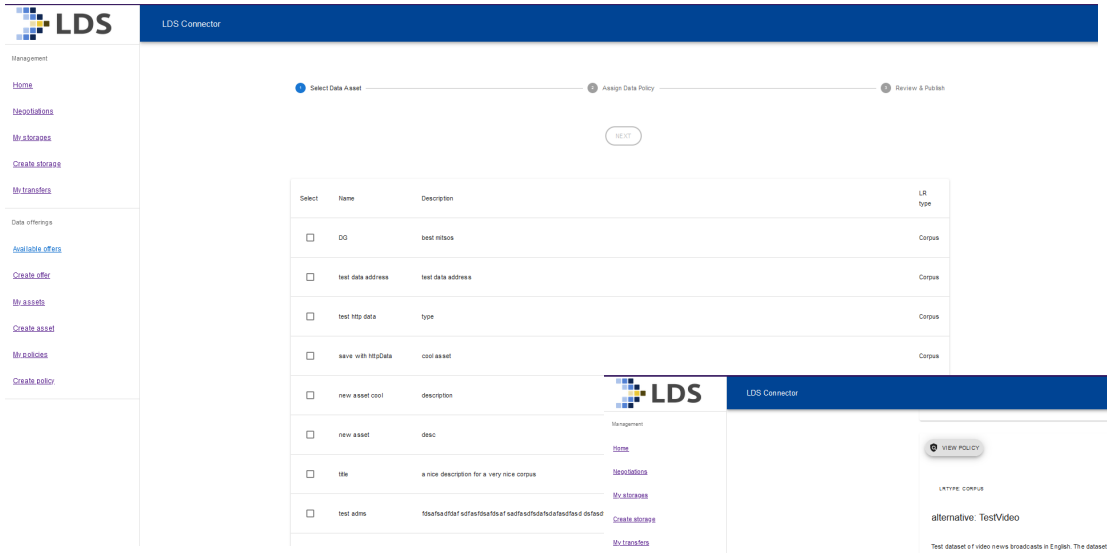


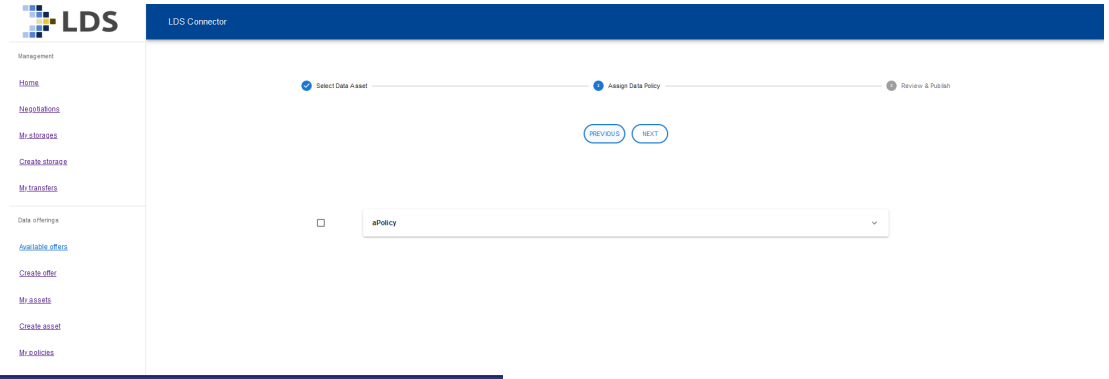


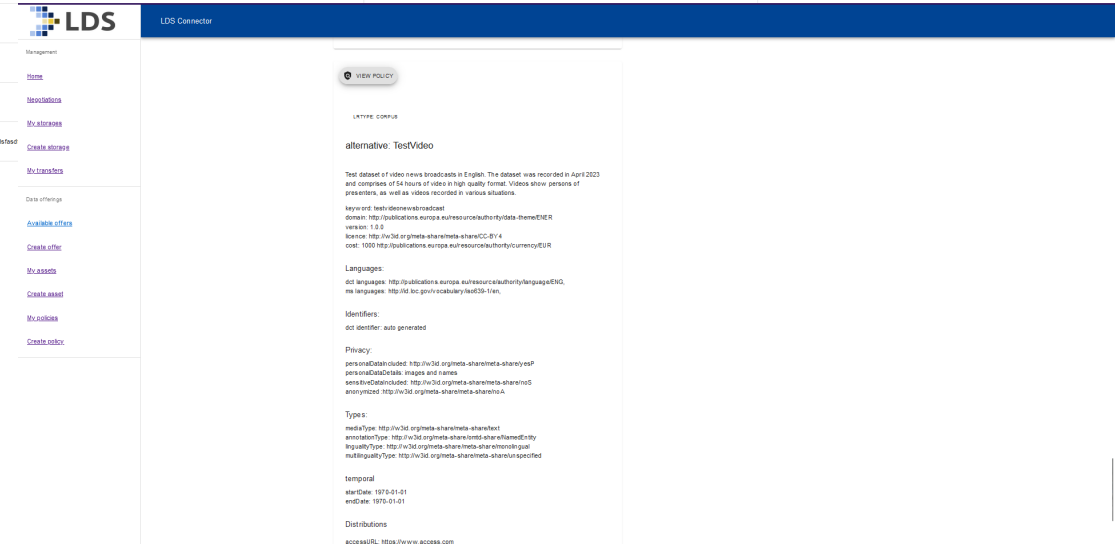


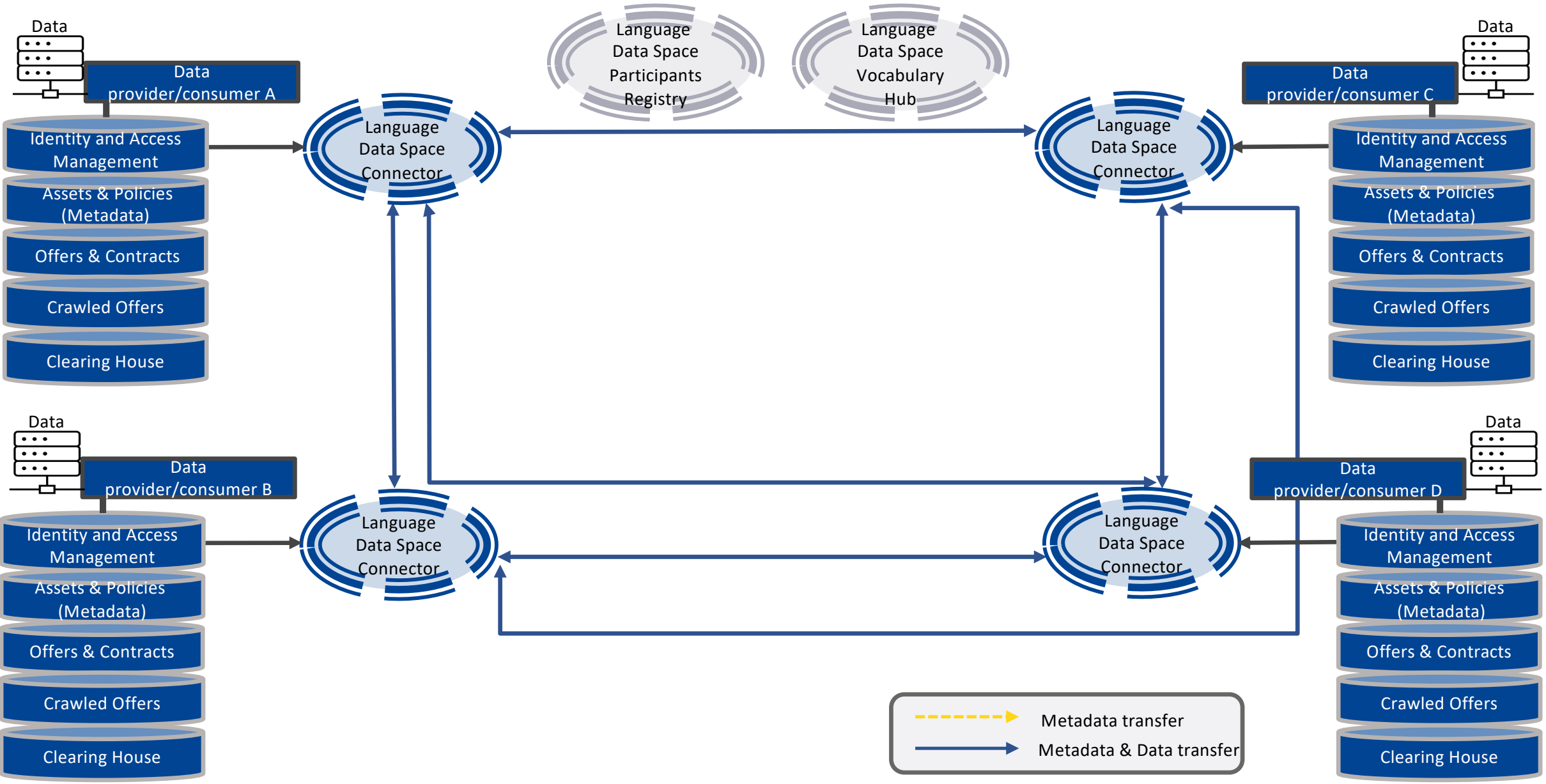

Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Language Data Space Connector
Language Data Space Participants Registry
Language Data Space Vocabulary Hub
Language Data Space Connector
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Language Data Space Connector
Language Data Space Connector
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts

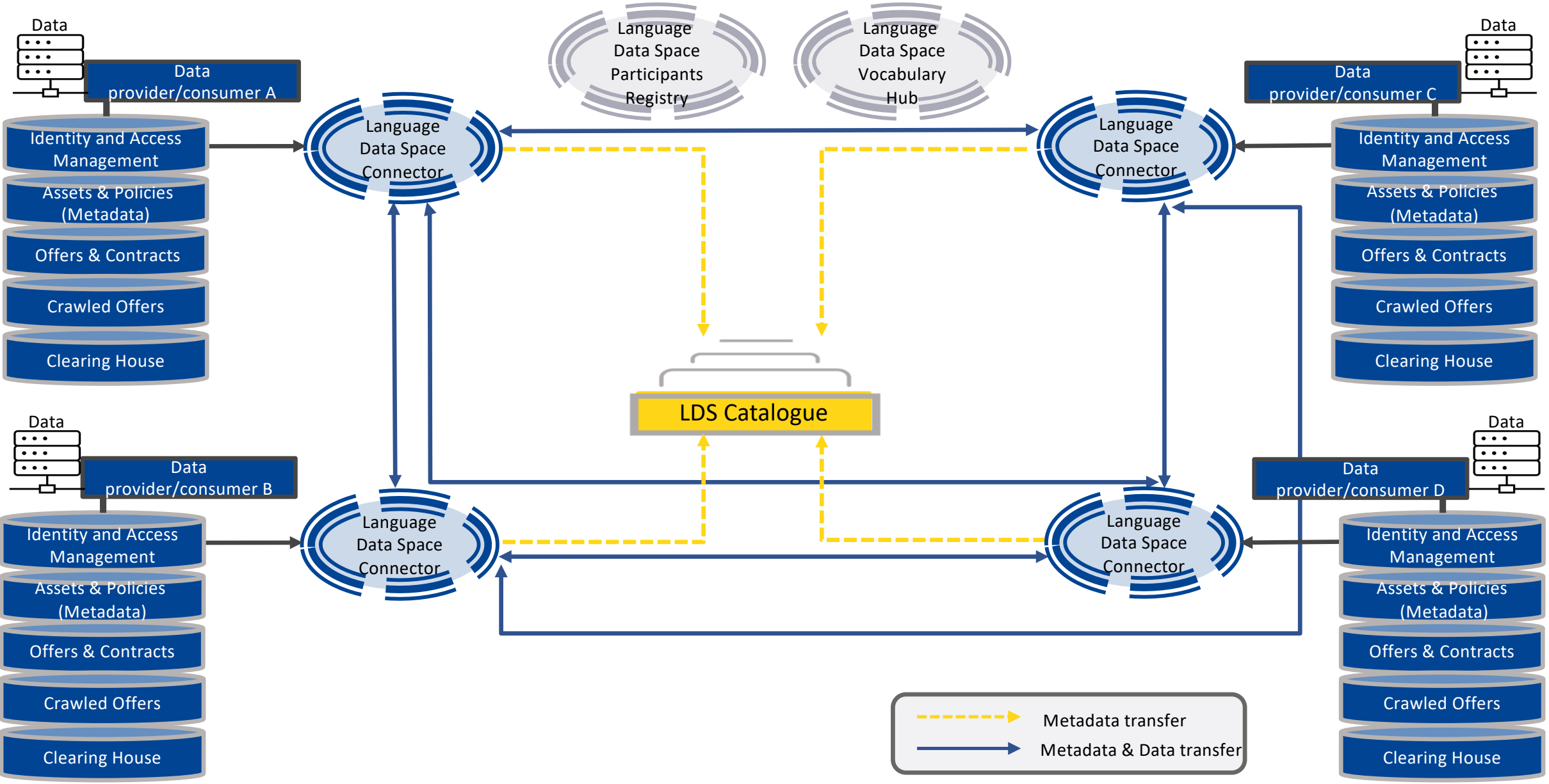
Creating Offers (Asset + Policy)

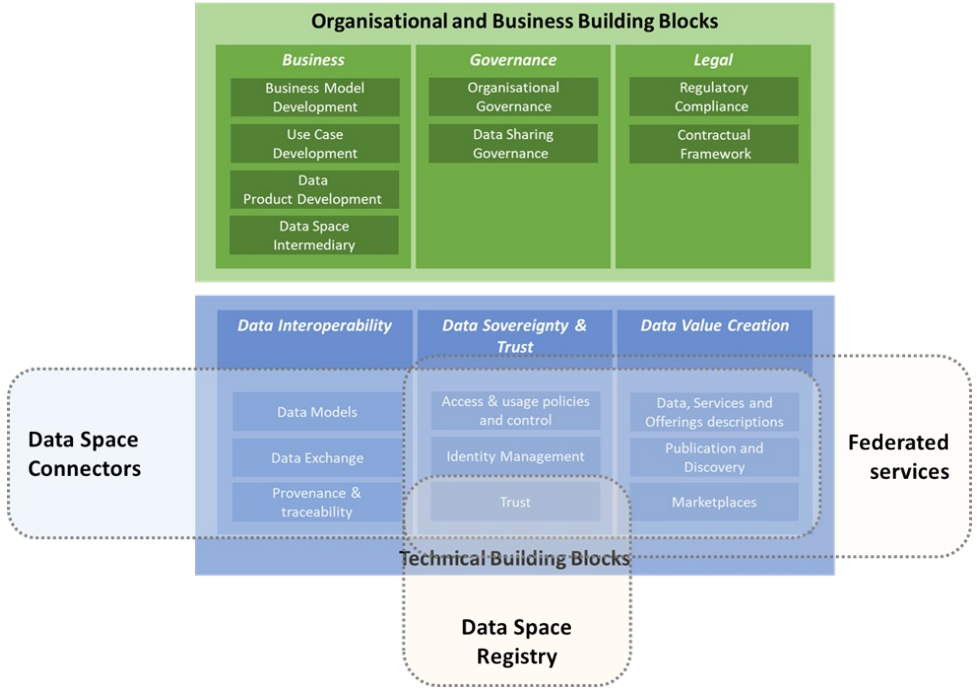






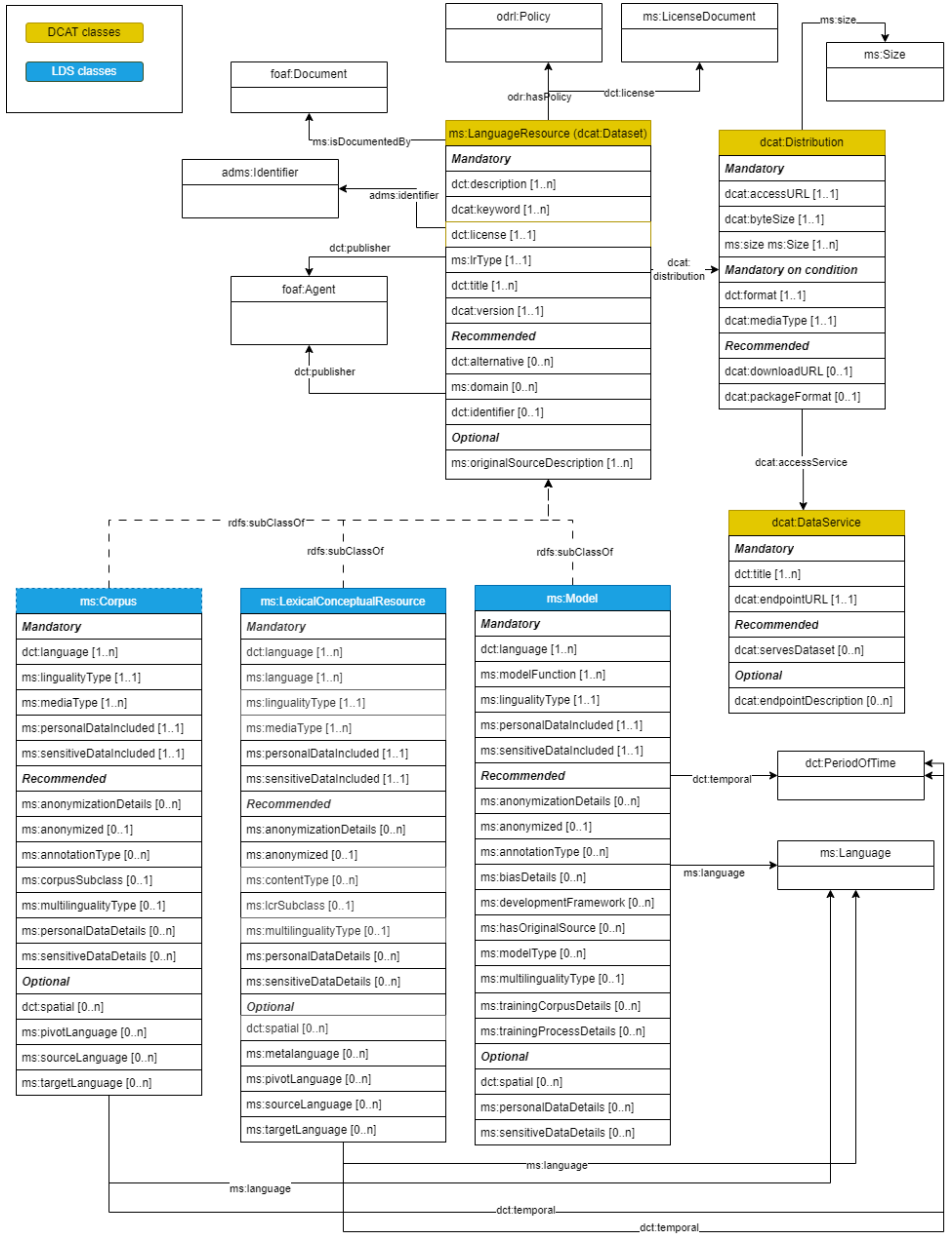

Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Language Data Space Participants Registry
Language Data Space Vocabulary Hub
Language Data Space Connector
Language Data Space Connector
Data
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Language Data Space Connector
Language Data Space Connector
Data
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Metadata transfer
Metadata & Data transfer


Language Data Space Participants Registry
Language Data Space Vocabulary Hub
Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Language Data Space Connector
Language Data Space Connector
Data
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Catalogue
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Language Data Space Connector
Language Data Space Connector
Data
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Metadata transfer
Metadata & Data transfer



Build on Existing Solutions

- Following the DSSC (see above) approach
- Eclipse Data Space Components (EDC)
- DCAT-AP, Language DCAT-AP (see right), ODRL
- Mappers from existing platforms



Next Steps

- Technical development, promotion, dissemination, governance etc.
- Collaboration with DSSC, Simpl and ALT-EDIC
- Collaboration with EU projects, e.g., HPLT, OpenWebSearch
- Collaboration with data spaces, especially Media and Cultural Heritage
- Collaboration with EuroHPC
- Adoption of LDS by industry and other organisations → grow the LDS User Group – https://language-data-space.ec.europa.eu
- Identify and make available new and fresh language data, especially from industry and covering all European languages and modalities

Common European Language Data Space

Thank you!

Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiļjevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitris Galanis, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdiņš, Elena Leitner, Valérie Mapelli, Hélène Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault and Leon Voukoutis

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

22/23/24-05-2024 LREC-COLING 2024
https://language-data-space.ec.europa.eu