EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY

Data
Accessibility, Multilingualism and Safer Internet

Language Data Space

Workshop Report

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

In line with the European Data Strategy and the launch of the DIGITAL Programme, the European Commission (CNECT.G3's Multilingualism Sector) organised a series of eight workshops. They targeted several business sectors (News, Broadcasting, Advertising, Publishing, Language Technology, Telecommunication Industries as well as Libraries, Archives and Public Administrations). The goal was not only to present the 'Language Data Space' concept, but also to gather insights from the different stakeholders' groups.

The Language Data Space aims to give stakeholders the opportunity to monetise their efforts in terms of language resources (data, tools, services, models, etc.), while also supporting the deployment of language models and language technology services for their businesses, in one single marketplace. Its objective is to create an interconnected and competitive European data economy for the valorisation and re-use of language resources.

The Language Data Space will be financed as follows:

- Framework Programme: Digital Europe Work Programme 2021-2022;
- Type of Action: PROCUREMENT;
- Indicative Budget: EUR 6 million;
- Indicative Time of the Call Opening: June – September 2022;
- Indicative Starting Date: Early 2023.

More than 100 attendees and their representative organisations (e.g., FEP, FEDMA, ETNO, EBU, etc.) participated in these workshops. All of them showed interest and identified several opportunities, for instance, monetising language data, counteracting the European Language Technologies landscape fragmentation and enriching it with high-quality data, covering different modalities, business domains and use cases. In addition, stakeholders maintained that indispensable 'enabling conditions' must be implemented and certain challenges have yet to be overcome, on a technical (e.g., promoting standards, normalising metadata, designing and developing the architecture), legal (e.g., complying with GDPR, implementing IPR clearance and correct licensing) or operational (e.g., defining governance, fostering sustainability and interoperability) level.

2. CONTEXT

The 'Language Data Space' or LDS is an EC initiative that aims to favour the exchange of language data and models across the European public and private sector. The LDS is designed to be a platform and a marketplace that co-ordinates the collection, creation, sharing and re-use of language data and resources and support the deployment of large language models and Artificial Intelligence (AI) Language Technology (LT) services. It will create new opportunities for European stakeholders to monetise their language resources and participate in innovative endeavours that can help scale their operations. The ultimate goal will be to allow the European LT industry to compete globally by building technological sovereignty and digital autonomy.

The LDS will be initiated as a procurement procedure within the DIGITAL Europe Work Programme 2021-2022 with an indicative budget of about EUR 6 million. The call for tenders is expected to be published between June and September 2022 and the procurement should start in early January 2023.

In order to raise awareness about the LDS initiative and gather insights, the European Commission (CNECT.G3's Multilingualism Sector) has recently organised a series of eight workshops targeting several business sectors. The list of the corresponding stakeholders' groups is reported in the table below. Organising separate workshops was deemed necessary to collect each group's specific requirements, but also to ensure sector empowerment in this endeavour.

	Stakeholders' Group	**Workshop Date**
1	News Agencies, Newspapers	26/04/2022
2	Broadcasting Industry	28/04/2022
3	Publishing Industry	04/05/2022
4	Advertising Industry	10/05/2022
5	Libraries, Archives and Public Administrations	12/05/2022
6	Telecommunications Companies, Call Centres	17/05/2022
7	Language Technology Industry	19/05/2022
8	ELRC National Anchor Points	01/06/2022

3. PARTICIPANTS

More than 100 attendees from all EU Member States covering several business sectors as well as their representative organisations participated in these workshops:

- From SMEs to large companies;
- Associations representing the above-mentioned business sectors, such as:
  - Members of EANA (The European Alliance of News Agencies) for the news;
  - FEP (Federation of European Publishers) for publishing;
  - FEDMA (Federation of European Data and Marketing) for advertising;
  - EBU (European Broadcasting Union) for broadcasting;
  - ETNO (European Telecommunication Network) for telecommunication;
  - LT-Innovate for LT industry.

Public administrations have been represented by several participants from the National Archives, National Libraries, European Language Resource Coordination National Anchor Points (ELRC NAPs), colleagues from the Publications Office of the European Union and stakeholders from previous relevant EC initiatives (ELRC, TAUS, etc.).

Other related activities have been represented by the corresponding units when appropriate:

- The Cultural Heritage Data Space (CNECT.G2)
- The Media Data Space (CNECT.G.2)
- The AI on-Demand Platform (CNECT. A.1)

For the full list of participants, please see 'Annex 1. List of Participants'.

Participants have been asked to present their ideas and insights about the LDS in **4 to 5 minutes** with a three-slide presentation articulated in three parts:

(1) Data management (data governance, architecture, design and modelling, integration, quality, security and metadata) and language models, tools and services used in their business;

(2) Opportunities which the LDS initiative can bring to their corresponding business;

(3) Possible challenges a business from the corresponding sector could face in adhering to the LDS initiative.

The remainder of this document enumerates the main key takeaways expressed by the participants during all the workshops terms of opportunities and challenges.

4. KEY TAKEAWAYS

Participants showed interest and identified several opportunities and challenges regarding the LDS.

4.1. Opportunities

All participants are in favour of the LDS and see it as an opportunity, especially in terms of:

Market

The LDS will provide several market opportunities by:

- Increasing the return from the investment in data collection and creation (e.g., through data re-use by other parties);
- Creating new business use cases, for example transitioning from being pure publishers to being data solution providers;
- Creating a new market for highly curated language data and/or high-quality annotated/labelled data;
- Enabling a virtuous circle from high-value language data to better AI services and the creation of new content and new language data.

Holistic Approach

The LDS platform will adopt a holistic approach:

- It plans to cover the whole spectrum of modalities, including text, speech, image, sign language(s);
- It will be a single platform for sharing/exchanging data and services;
- It will enable enriching available data for low-resource languages;

- It will give access to industry-specific data, e.g., industry-specific terminology and its standardisation.

Services

The LDS will:

- Provide legal guidance;
- Provide standardisation;
- Provide evaluation metrics for data and model quality measurements;
- Support data monetisation.

Networking

In terms of networking, the LDS will:

- Be a 'unifying' initiative, as it will reduce the LT landscape fragmentation;
- Allow 'bringing together' interested parties, also in case of actors whose core business is not purely LT-related;
- Facilitate finding partners to develop language models;
- Bridge the cultural and linguistic digital divide;
- Increase data availability, across domains and languages, for training systems, and provide a better overview of available data;
- Facilitate market insight, business prospection by better aligning:
  1) the offer of data, tools, services and
  2) needs of potential users in public administrations and industry.
- Enable the easy identification of new types of tools and services;
- Help SMEs to increase their visibility on the LT market.

4.2. Challenges

Even though all participants are in favour of the LDS and showed interest in being on board, they also considered that indispensable 'enabling conditions' must be met and some barriers removed.

In particular:

Technical barriers

Participants showed interest in the 'second use' of their data (i.e., use the data to build NLP tools like Machine Translation, Automatic Speech Recognition, Language Models, etc.) but insisted that this should not interfere with their core business. Technical

solutions should be provided to overcome this risk, for example through the availability of tools to change format or shuffle sentences. An example of that is using audiobooks to build algorithms for Automatic Speech Recognition. Audiobook publishers need a guarantee that the data they are providing/selling is transformed so that it cannot be used to reconstruct audiobooks. This could be seen as an additional technical guarantee to the adequate licencing framework covering this 'second use' of the data.

Other technical challenges have been also raised, in particular regarding:

- Distributed data management, including rights data management;
- Distributed data governance;
- Metadata standards needed to cope with the multitude of structures and formats (e.g., sign language);
- Metadata quality, completeness and consistency.

Legal barriers

Several legal challenges, especially those connected to IPR and GDPR compliancy, have been mentioned as major obstacles:

- Importance of IPR clearance;
- Data security: availability of anonymisation of personal identifiable information and its legal validity;
- Importance of finding the right licencing solutions to let data provider exercise their right of full ownership of their data;
- Plethora of access right statements/legal conditions, access scenarios;
- Fear that LDS will be another legal constraint (like IPR, GDPR) and not an open, enabling solution;
- Privacy preservation in AI as a Service;
- Proprietary knowledge preservation.

'Access' barriers

Some participants hinted that the LDS should remain easily accessible in terms of both interface and tools, and its realisation should be simple, focused and user-friendly. Small businesses are confronted with limited skills resources. They need help, at least in the beginning, in redressing possible gaps to have the opportunity of availing themselves of the project results. LDS should avoid a heavy bureaucratic loads associated with EU projects.

Governance barriers

The governance mechanism has been qualified by some participants as complex to understand ('The leading question will be: who is in the driving seat?'). They expressed that the focus needs to lie on the industry, supporting it.

Interoperability challenges

Multiple participants pointed out that interoperability will be an important aspect to be taken into account at different levels:

- Interoperability in terms of data formatting and tools compatibility;
- Semantic interoperability and subsequent findability across languages;
- Semantically linked data, hence a need for topic filtering;
- Interoperability with the other Data Spaces.

Neutrality challenges

It was emphasised that the LDS should remain neutral to:

- Content type: as the LDS will have its own ways of collecting and preparing data, the diversity of stakeholders' data could be lost;
- Business model: the LDS blueprint should enable different business models from different stakeholders' groups as each group might have very specific needs.

Financial challenges

Some participants brought up the risk that the LDS is underfunded.

Market challenges

Some market challenges have also been raised:

- The LDS should act as an aggregator and not as a market distorter. If squarely geared, it might interfere with the current market players, their investments and businesses, when, actually, it should be supporting them in moving faster and in a more resilient way, not interfering on the market;
- The LDS should promote balance between data protection and market functioning. The intention is to create this potential in Europe (before EU industries go and use models trained in the US, on American data and under American legislation).

Operational challenges

The operational capacity of the LDS will be key in terms of:

- Structure in charge;
- Business models;
- Long-term sustainability;
- Market segments.

Strategic challenges

The ultimate goal of the LDS is to help the European industry compete globally. It is therefore important to strategically match American services in the LT and AI field. Another strategic aspect is to build trust throughout the data sharing process among all actors (public administrations, media, publishers, etc.).

Ethical challenges

The challenges around AI in general and LT in particular have been also pointed out from an ethical point of view. The LDS should take account of the concept of 'Responsible AI' in its governance framework.

Benchmarking challenges

Several participants highlighted the need for real-life high-quality data for evaluation and benchmarking. This implies:

- Defining common standards and guidelines for data quality;
- Making specialised benchmarking available;
- Providing tools and services to ensure data quality, e.g., data identification, deduplication, versioning, persistence, de-identification of personal data;Addressing the unbalance in language coverage (low-resource languages require various techniques, e.g., transfer learning and innovative approaches).

4.3. Threats

The following aspects have been identified as threats by some participants:

- 'Academic' outcome good for research, but distant from the market;
- Limitation of data to specific domains and sectors;
- Low-quality language data with inconsistent content;
- Unfair leverage of LDS by the global tech giants.

4.4. Additional Findings and Stakeholders' Contributions

It is worth mentioning that one of the findings of the workshops is the TDM protocol, a text and data mining protocol (a standard based on W3C to have machine-readable information) to help publishers promote the commercial usage of their content at fair terms. It was presented by the Federation of European Publishers (FEP). The Publishing Industry developed this protocol to technically implement Art 4 of Directive 2019/790, cf. Text and Data Mining Reservation Protocol Community Group (w3.org).

5. ANNEX 1. LIST OF PARTICIPANTS

BUSINESS SECTOR	COMPANY / ORGANISATION
ADVERTISING INDUSTRY	FEDMA – Federation of European Direct and Interactive Marketing
	IAB Europe – Interactive Advertising Bureau
	Kellenfol Advertising S.L.
	Talpa Network
	UNA COM – Aziende delle Comunicazioni Unite
	VIA NEDERLAND
BROADCASTING INDUSTRY	AER – Association of European Radios, VAUNET – German Media Association
	Deutsche Welle
	EBU – European Broadcasting Union
	RTBF – Radio Télévision Belge Francophone
	Star Channel, Dromos 89.8fm
	VAUNET – German Media Association
LIBRARIES, ARCHIVES AND PUBLIC ADMINSTRATIONS	BVO – Austrian Library Association
	National Archives of Hungary
	National Central Library of Florence
LT INDUSTRY	ADAPT Centre
	Athena Research Center
	AX Semantics
	Business Innovation Consulting
	Coreon GmbH and ESTeam AB
	Cortical.io
	CrossLang NV
	DFKI – German Research Center for Artificial Intelligence
	ELDA – Evaluation and Language Resources Distribution Agency
	EMF Services
	Expert.AI
	GIRAF PM e.K.
	Hungarian Research Centre for Linguistics
	Hensoldt
	Intersystems
	SDI Media Latvia SIA
	Le Voice Lab / Linagora
	Limecraft
	Linguist, Conference Interpreter
	Linguistic Systems
	LT Innovate
	Mensource

	Mozajka
	Neticle
	Neurolingo
	Orco
	Pangeanic
	Phonexia
	Sign Time
	Sogedes
	TAUS
	Text United
	TILDE
	TMServe
	Université de la Sorbonne Nouvelle
	WebLyzard technology
NEWS INDUSTRY	AFP – Agence France-Presse
	DPA – German Press Agency
	PAP – Polish Press Agency
PUBLISHING INDUSTRY	Elsevier, RELX Group
	ePublishers
	FEP – Federation of European Publishers, AIE - Associazione Italiana Editori
	Slovak Print and Digital Media Association
TELE-COMMUNICATIONS INDUSTRY	ETNO – European Telecommunications Network Operators
	GBA Call Centers
	Orange
	Telefonica
	TELENOR ASA
	Thales Group
	VODAFONE Business

6. ANNEX 2. WORKSHOP AGENDA

The workshop will bring together key stakeholders from the Language Technology industry to gather their ideas inputs for the upcoming 'Common European Language Data Space (LDS)' endeavour, set in the DIGITAL Work Programme 2021-2022.
Discussions will mainly revolve around

a) Data management in the Language Technology industry
b) Connections between businesses and the Language Data Space

– AGENDA –

8.50 – 9.00	**Opening of the Webex conference room**	EC
9.00 – 9.30	**Welcome and Scene-Setting** • Introduction • Flow of the day	Philippe Gelin

9.30 – 10.55	**Tour de table** *(max. 4/5 minutes per participant)*
	• Introduction
	• Data Management
	• LDS-driven opportunities for businesses
	• LDS-related challenges for businesses
	Discussion
	• Business-Language Data Space Connection(s)
	  • Benefits and risks
	  • Advantages and disadvantages
	  • …
		Stakeholders / Plenary

10.55 – 11.00	**Conclusions and wrap-up**	EC

7. ANNEX 3. USEFUL LINKS

TED (the online version of the 'Supplement to the Official Journal' of the EU where the procurement will be published)

European Data Strategy

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data

Digital Europe Work Programme 2021- 2022 and DIGITAL on Funding and Tender Opportunity Portal

Staff Working Document on Data Spaces

European Language Resource Coordination (ELRC)

European Language Grid (ELG)

Data Governance Act (for more information on the EDIB)

EDIHs (for more information on European Digital Innovation Hubs)

EuroHPC (for more information of High Performance Computing)