data.europa academy webinar ‘Data spaces: experience from the European Language Data Space’ - 11/04/2025 - Q&A

Question	Answer
You mentioned that there are significant efforts in the EU to develop large language models similar to GPT-4, with Mistral being one such example. However, many EU citizens still seem to prefer using models and tools like GPT-4 or DeepSeek over these European alternatives. Why do you think that is, and what can be done to encourage greater adoption of EU-developed AI tools?	Cfr.: https://openeurollm.eu/ I believe that OpenAI and other US players took large legal risks to arrive first on the market and deploy their solution. There is no doubt that they took benefit from the empty space, but their solution remains shady from a legal point of view (from both Copyright and GDPR). EU alternatives are compliant with the regulation, but I believe that the main attractiveness resides in their own European creation. GPTs are so complex that beyond the various languages, they are embedding cultural aspects from the training sets. And it is important that EU users (personal and/or companies) stay close to their market. EU LLMs developers pay great care to ensure that such “cultural/language bias” is incorporated. There is also, of course, the user’s data collection aspect. LLMs applications such as chatGPT are hungry for personal data, not only to better answer the users but also to refine their models. These tools are more sensitive to personal data and, as it stands, for non-EU applications, these data are crossing the ocean and re-used for unknown purposes.
Are there audit trails or monitoring tools currently used to track data lineage and reuse across the LDS infrastructure?	As explained, the LDS infrastructure will not analyse nor take responsibility for the Data transferred within the infrastructure.
Should the responsibility for monitoring and enforcing license compliance rest primarily with the operators of open data portals?	As explained, the LDS infrastructure will not analyse nor take responsibility for the Data transferred within the infrastructure.
How do LDSs handle data provenance—specifically, what frameworks or methodologies are being used to track the origin, versioning, and individual contributions to language datasets, particularly when these are collaboratively developed or aggregated from multiple parties? In such cases, how can the definition and documentation of a dataset’s origins be structured to fairly reflect and safeguard the	LDS does not monitor or take any responsibility for the quality and provenance of the datasets shared by providers. The resource description (aka. metadata record) includes elements that providers can and it is recommended to use to document relevant information; it’s the providers’ responsibility to appropriately fill them in.

rights, roles, and interests of all contributing stakeholders?	
Where is the LDS data model documented?	The LDS model (LanguageDCAT-AP) is in the process of being documented; a link to the documentation will soon be provided via the LDS official site (https://language-data-space.ec.europa.eu/index_en)
What is the procedure to update a certain data offering in the LDS context? Which aspects should we consider when a data offering is removed or eliminated on signed contracts?	According to the LDS data lifecycle, providers can withdraw their offers from their (local) catalogues and add new offers for the same dataset with a new policy. The withdrawn offers remain available only to those consumers that still have active signed contracts (e.g., purchased a dataset but haven't yet downloaded them) until the contract terms are fulfilled. If a provider fails to fulfill their contract, the consumer can file a complaint with the LDS authority (LDS Governance Board).
Can you share a concrete example of how the LDS platform can be used in a real-life scenario	The LDS platform can be used by publishers, media companies, etc. that wish to share/exchange/monetise their datasets and make them available, for instance, for training LLMs. Developers of LLMs that wish to train new models or refine existing models with domain-specific data can use the LDS marketplace to discover new datasets and obtain them in compliance with their policies (licensing terms). The videos presented show the actual use (representing the current stage of development).
Considering that audiovisual data is personal data, how can data providers comply with GDPR obligations when sharing the data to a data consumer? i.e. transparency obligations, data subjects requests that can oppose to their data being shared to another party, (how to even comply with right of erasure?). Does LDS helpdesk support this compliance?	This obligation is part of the Data owner and under their own responsibility. The LDS helpdesk can provide guidance, but will not take legal responsibility. One possibility is for the data owner to keep control of their data and remove the information of the data subject.
Will the LDS include specific terminologies, for example sets of translations of medical terms?	The infrastructure is compatible with such data.
How LDS will be related to EOSC (European Open Science Cloud)?	The EOSC EU Node can be conceived of as a Data Space for Open Science, while the Language Data Space will offer a marketplace for language resources (language datasets and language models). Collaboration between the two, as well as across other/all data spaces, is part of the vision and can be established at later stages.

What is the status of language dcat ap?	The LDS model (LanguageDCAT-AP) is in the process of being documented; a link to the documentation will soon be provided via the LDS official site (https://language-data-space.ec.europa.eu/index_en)
Which other controlled vocabularies are used beyond the list of languages?	The LDS model (LanguageDCAT-AP) uses various controlled vocabularies depending on the data property, e.g., countries, scripts, file formats, domains, etc. These will be made publicly available together with the model.
What is the difference between LDS and sharing language datasets-distributions as DGA or open-data datasets-distributions?	The European Language Data Space aims at building a trustworthy and effective data market for the exchange of language resources in the public and – even more importantly – in the private sector, in line with the EU Data Strategy. Open datasets can certainly be shared through LDS following the LDS workflows.
The data space uses open standards such as DCAT-AP which is very good. Is the catalogue code for example available as open source as well?	The LDS software code will be made available with an open source licence; during the development phase, the code repositories remain private.
Has the CdT been approached to share its translation memories?	Not Yet. This is an excellent reminder.
Are you planning the LDS connection with Simpl?	LDS should in principle be compatible with Simpl. If the LDS and Simpl roadmaps/timeplans allow it, LDS intends to adopt generic Simpl modules and adapt them to LDS requirements.