Open Future submission to the public consultation on the European Data Union Strategy

We welcome the opportunity to contribute to the Commission's consultation on the Data Union Strategy. We support the Strategy's goals of strengthening Europe's data ecosystem by improving interoperability, data availability, and access to high-quality data for innovators.

We appreciate that the input paper circulated during consultations of the Strategy refers to the needs of businesses and society. Creating conditions for trusted and fair sharing and use of data cannot be understood only in terms of creating a single market for data and supporting the European economy. And the innovators, whom the Strategy is to support, are not just business actors.

The European data ecosystem should therefore be considered a key element of the European Digital Public Space, to which the European Declaration on Digital Rights and Principles refers. Control and governance of data are not only vital for driving innovation and competitive advantage but also for safeguarding rights, and ensuring both economic and social sovereignty.

European Data Commons

In order to achieve these goals–in particular those related to supporting data availability, access and use–one of the goals of the strategy should be the creation of a European data commons: a pool of data that are collectively stewarded and governed.

Enrico Letta puts forward a similar proposal in his "Much More Than a Market" report, where he positions the European Knowledge Commons as a central pillar of a strategy to foster collective intelligence and drive innovation. He envisions it as "a centralised, digital platform providing access to publicly funded research, data sets, and educational resources." This empowers citizens, researchers, and businesses alike, allowing them to tap into a wealth of knowledge for innovation and societal progress". However, such a data commons does not need to take the form of a centralized platform - instead, it should be an ecosystem of decentralized solutions, including the data spaces and data labs, and also data sharing infrastructures and data sources.

Building a European data commons requires public infrastructure that can be consumed non-rivalrously, by actors from the public, private and civic sectors. Such a data commons is particularly suited to address the needs of AI development, and other machine learning uses, which benefit from the pooling of data into large scale datasets.

The data commons approach entails a governance framework that balances public interest objectives, economic value creation, and fundamental rights protection. In other words, such a framework ensures access to data at a proper level of openness, while also ensuring democratic control and enabling public value to be generated. For a data commons to exist, an institutional layer is necessary, with entities that enable sharing, ensure public value generation, and facilitate participatory governance.

The European data commons needs to encompass both Open Data and data that cannot be shared openly and is made available through gated mechanisms. The general rule should be that as much data is shared as possible, with as many restrictions as necessary. It is important to define the essential restrictions on openness, taking into account the interests that must be protected, such as, for example, privacy, creators' rights. In governing any data set, it is also essential to map benefits and aspects that are improved by making the data set open and available, and the challenges that might arise.

A purpose-driven approach

The Data Union Strategy should have a strategic focus that steers the development of the data commons. In other words, the Strategy cannot adopt a "more data is better" approach that easily leads to extractive processes and to the risk of developing, based on this data, technological infrastructures without a clear purpose.

Use of data, in particular for applied AI solutions, needs to happen in a purposeful manner, especially in sensitive sectors like public education or health. There are few prior examples of evidence of the positive impact of AI on these spheres of life, and data-driven deployment of AI cannot be guided just by a vision of productivity gains. By taking a more discerning, realistic view of emerging technologies, EU funding could avoid techno-solutionism and instead prioritize initiatives that meet genuine needs and produce tangible value.

While the Data Union Strategy is primarily an industrial policy instrument, it should be guided by a dual purpose: strengthening Europe's AI and data economy, while also ensuring that data infrastructures serve the broader public interest. This means addressing the needs not only of AI developers, but also of European citizens, communities, and public institutions. This also means that supporting the development of AI technologies cannot be the sole goal against

which data access measures are evaluated. Greater data availability should serve various other purposes, including those that do not relate to machine learning.

Data commons, fundamental rights and sustainability

By recognizing and addressing the tension between the sharing of resources and the individual and collective rights related to these resources, a commons-based approach intends to preserve the status of resources as public goods while mitigating negative consequences that might arise from sharing. Potential issues arising from the unrestricted sharing of online resources include privacy violations, concerns about content creators' rights, and the fair treatment of contributors involved at various stages of data curation, as well as the sustainable and just maintenance of digital commons as such.

Regulatory safeguards should not be lowered under the guise of simplification, and data sharing initiatives should not bypass consent and accountability mechanisms. Furthermore, publicly funded initiatives like the AI Factories (and the planned Data Labs) should lead by example and uphold strong safeguards. This means designing and governing AI development in line with fundamental rights principles.

We do not see a need to re-examine or revise the GDPR, which already allows AI training on personal data where lawful and justified. If anything, the large-scale use of public data in AI training should be seen as a reason to reinforce the protections offered by the GDPR.

Finally, the Strategy needs to take into consideration concerns regarding environmental justice and sustainability of AI development. Data Labs, and the data centers at the heart of AI Factories and Gigafactories, need to operate within planetary boundaries. This entails embedding accountability, transparency, and climate alignment into the entire lifecycle of data centers and computational resources. Publicly funded data infrastructures should lead by examples in measuring and mitigating environmental impact.

A data pathway to AI development

Investments in computing infrastructure for AI development and computing power in particular (including the already established AI Factories, and the planned Gigafactories) are not enough. The AI Continent Action Plan correctly identifies access to data as essential for the success of European AI development.

The Data Labs, proposed in the Action Plan, are needed to fulfill the goals related to the availability of data for AI development. These Data Labs should function as part of a larger

framework provided by the European Data Commons, as dedicated entities responsible for data federation, data quality and access to the data, in the space of AI development. These Data Labs also need to have a bridging function, between the AI development sector and the entities stewarding various data sources and collections.

Lack of data that is available for lawful use, of sufficient quality and clear provenance is a bottleneck for EU AI developers that is as crucial as lack of access to computing infrastructure. There is a clear need for public data that complies with the EU regulatory framework and can be used in public and private AI systems developed in the AI Factories. However, there is a risk that AI Factories will support projects and entities with proprietary approaches to dataset development. Steps need to be taken to ensure that resources are directed toward building public datasets and data repositories that serve the public interest AI ecosystem. The European Data Commons framework would ensure that ethical and legal data sharing is at the heart of the AI Continent effort.

Furthermore, the Data Labs should be part of an open ecosystem. The data sharing infrastructure, processes and pipelines should be based on principles of openness and interoperability, with as many tools as possible shared openly; and with interoperability mechanisms in place, to avoid lock-in.

Establishing a European Data Commons as part of the AI Continent Action Plan requires greater investments in:

- Infrastructure for storing and making available data held by public institutions, and potentially also for digitization efforts;
- Developing data refinement and management pipelines and toolkits that improve the quality of data, with a focus on making it useful for AI training;
- Creating and releasing specific, high-value datasets.

Data Infrastructure for Generative AI Training

The Strategy should establish and support the development of a dedicated infrastructure for generative AI training that exists alongside data spaces, and plays a key role in the European Data Commons. European Union needs this data infrastructure to support the development of European, public alternatives to models owned by foreign companies.

There is a need to build in Europe AI models created and deployed with as minimal dependencies on private infrastructures and solutions as possible. These models should be

digital public goods: open source technologies that are democratically governed and designed with the purpose of addressing key social challenges. The creation of an open source large AI model for all European Union languages should therefore be one of the missions of European Union's AI policy - supported by measures related to

As general-purpose technologies, these models can provide the foundations for a wide range of technological applications, including those aimed at addressing various societal challenges and sectoral needs. And without such models, AI development runs the risk of creating more dependencies and of entrenching concentrations of power, with adverse impact on competitiveness and innovation, sovereignty and sustainable growth.

The intent to train such models is expressed in the AI Continent Action Plan, and a range of existing initiatives, such as the OpenEuroLLM consortium, are aimed at developing them. There is also a range of companies and research initiatives building today European open source models of various sizes and capacities. Still, there is a need to build more robust and competitive models that can serve as "capstones" for an ecosystem of data-driven services and solutions.

We therefore propose to build a data infrastructure that focuses on the specific goal of developing state of the art, public AI models for European languages.

The creation of such AI models requires such a cross-sectoral approach, as useful pre-training data can be found in various of these sectors. It also requires aggregating data from various languages and regions of the European Union. At the same time, existing data spaces – such as the heritage and language data spaces – are of key importance for this effort.

To this end, the mission of Data Labs needs to expand beyond aggregation and data pooling. Lack of publicly available AI training datasets means that they need to be intentionally created. The new Data Union Strategy needs to ensure not just mechanisms for aggregating and sharing data, but the existence of these datasets. This requires in particular working with various organizations that are stewards of collections, including academic and research institutions, broadcasters, archives and other heritage institutions. Support should focus both on making more data publicly available, and on refining data into collections and AI training datasets. This means investing in infrastructures for aggregating data, as well as in institutional infrastructures that will enable public institutions to share their digital collections. Support and investment is also required for digitizing collections and obtaining licenses.

The Data Labs, and the generative AI training infrastructure, need to not only cater to the demand of companies and organizations developing AI models and systems, but also center

the interests of these various stewards of collections. They also need to ensure participation of citizens in the governance of datasets, so that collection, access and reuse of data for AI development is driven with a focus on the needs of European citizens and communities.

This vision for developing public AI infrastructure, centered around the development of a capstone model and supported by data and compute investments, is developed in more detail in the "White paper for public AI", published by Bertelsmann Stiftung in partnership with our foundation.

Conclusions

The objective of the Data Union Strategy should not be to amass as much data as possible, but rather to provide access to data that is genuinely useful, to ensure its fair and ethical use, and align it with specific societal goals. The European Data Commons must ensure that data collection and use are guided by clear, publicly beneficial objectives and purpose-driven use cases. The development of public AI models for European languages is one such objective. Others include data-driven solutions to challenges in critical areas such as health, sustainable development or climate preparedness.

Building a European Data Commons will contribute to the Data Union Strategy's goals of increasing the use of data in a safe manner, empowering trustworthy AI and enhancing Europe's competitiveness. Such a data commons will also foster public interest use of data.

In building a European Data Commons, Europe will create infrastructures and solutions that support public interest innovation, safeguard fundamental rights, and ensure that diverse communities and regions benefit from Europe's data. Centering shared resources, civic participation, and the public good will support a vision of sovereign, European data infrastructure.