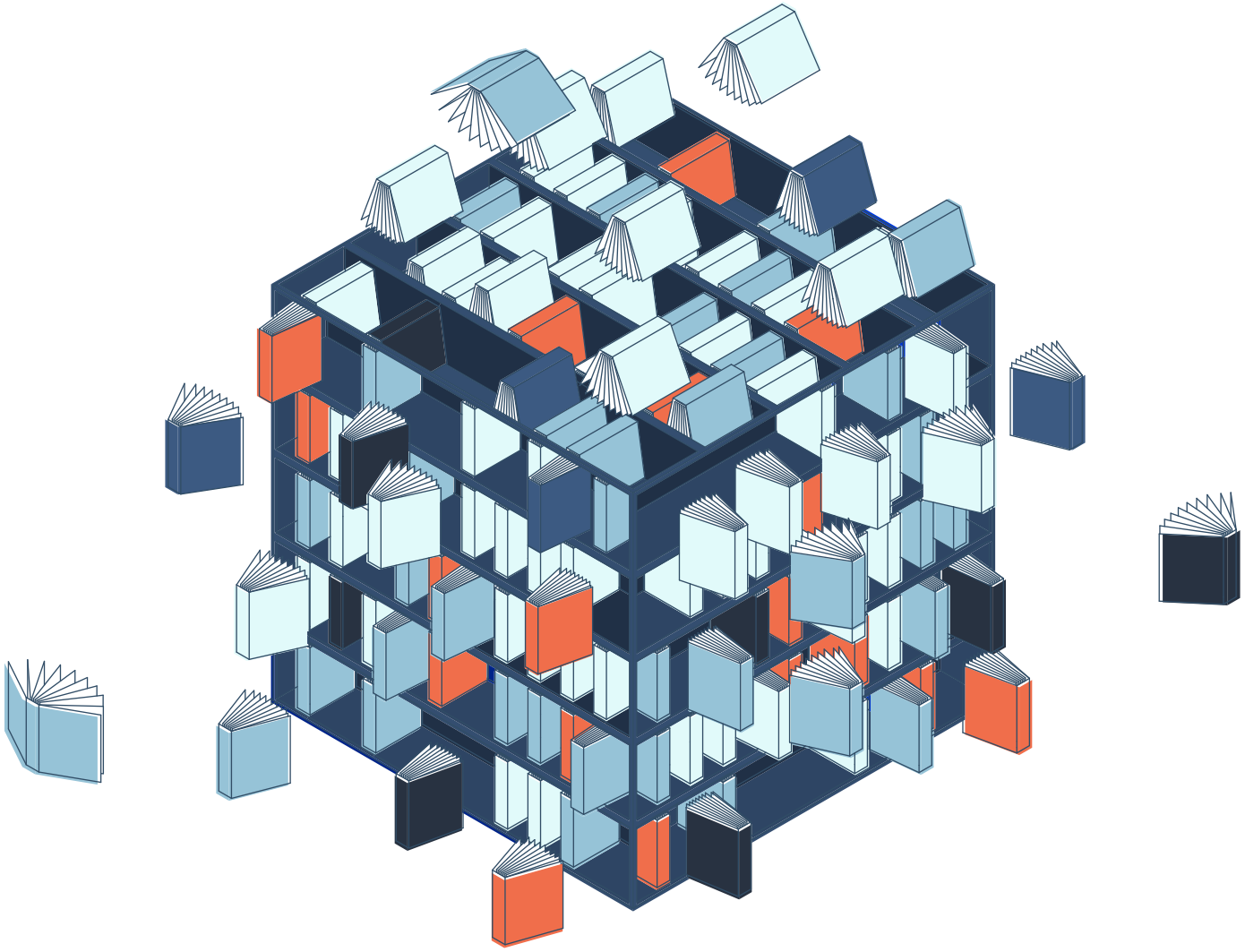
Outline for a ***European Books Data Commons***



OPEN _FUTURE

europeana

November 2025

Introduction

The idea for a European Books Data Commons emerged in reaction to the publication of the Open Future / Glushko & Samuelson Information Law and Policy (ILP) Lab report on Demonopolizing the European Public Domain in December 2024. It also builds on previous work done by Open Future, Proteus Strategies, and Creative Commons on developing a Books Data Commons for AI training.

In the first half of 2025, Open Future and the Europeana Foundation organized a series of structured conversations with European libraries and other stakeholders to validate the overall idea and to generate input on how to put the idea into practice. In this context, we also engaged with the Institutional Data Initiative at Harvard to exchange knowledge and good practices, as the Initiative is in the process of implementing a similar approach in the US. The concept outlined in this document has been developed based on insights obtained during these interactions.

Background and problem statement

The idea of a European Books Data Commons builds on the observation that, as a result of its digitization partnerships with a number of European libraries (hereafter "Google Books partnerships"), Google currently has unparalleled access to a large collection of digitized versions of public domain books from their collections. The Google Books project's structure—where Google has bilateral agreements with each library partner—means that Google is the only entity with structured access to the full collection. Each individual library can only access the digital versions of the books it provided. It is important to note that there is nothing inherently wrong with such an arrangement[1]: the agreements established that Google digitizes the books in return for making digital copies available to the providing library, with restrictions on how libraries could make these copies available for bulk access and reuse. This enabled large-scale digitization that otherwise might not have been possible. At the same time, the structure of the agreements has resulted in fragmented public access across different libraries and varying levels of availability in their online offerings.

With renewed focus on digitized books as training data for large language models (LLMs), library community members are exploring a single access mechanism to bring together digitized versions from partner libraries—one optimized specifically for AI training purposes. This represents a shift from the earlier paradigm of digitization, which emphasized discovery and access to individual works, toward a new requirement for aggregated collections that can serve as high-quality datasets for computational uses.

It is this type of access mechanism that we refer to as the European Books Data Commons (EBDC). The EBDC would provide a large, high-quality dataset of book scans—both as images and as full text—freely available for various uses, including training AI models. It would be developed and managed as a commons through the joint efforts of multiple libraries, with

[1] Although some of the early agreements contain exclusivity clauses (up to 15 years) that exceed the maximum term allowed under the 2019 Open Data Directive. See also: https://openfuture.eu/publication/demonopolizing-the-european-public-domain/

collective governance mechanisms in place, and means to ensure financial and environmental sustainability. The dataset would both facilitate access to books as training data and set appropriate rules for such uses.

In doing so, the EBDC would address the demand for ever larger amounts of high-quality training data while consolidating costs for development and infrastructure on the supply side. It would also create a more structural backup of the digital collections that have resulted from the Google Books partnerships. This is important since many of the library partners partially rely on Google as the repository for digitized copies.

Creating a European Books Data Commons would align with several existing policy initiatives. It could become part of the common European data space for cultural heritage and would align with the objective to increase the availability of high-quality datasets for AI training that will underpin the European Commission's upcoming Data Union Strategy as well as a wider set of policies that advance the EU's ambition to become an 'AI continent'.

The remainder of this document will sketch out how we envisage the European Books Data Commons working in practice. This is an idealized sketch that may need adjusting depending on the context in which it is implemented.

Purpose of the EBDC

At its core, the European Books Data Commons is a piece of public digital infrastructure that exists to provide access to large, high-quality datasets of book scans and full text. These datasets are freely available for various uses, including in forms optimized for training AI models. The European Books Data Commons will provide a centralized interface for accessing the data, which can be stored either centrally or in a distributed fashion.

As part of this, the EBDC will engage in data cleaning, data labelling, and other forms of data processing, but it will not engage in data curation. Data curation occurs either at the level of the data provider (libraries) or at the level of EBDC users.

Problems addressed by the EBDC

Once implemented, we would expect the EBDC to contribute to addressing the following problems:

- The EBDC should allow libraries to store their Google-digitized collections on public infrastructure under their control. This infrastructure should provide clear provenance information to ensure data authenticity and quality.
- The EBDC should support libraries in optimizing the data they have obtained from Google for various types of uses, including—but not limited—to individual access by humans and bulk access in the context of AI training.
- The EBDC should provide researchers and AI developers with a centralized point of access to European digitized public domain books. A centralized access mechanism can also reduce excessive and/or redundant data scraping.

- Bringing together the collections from a variety of library partners via a central point of access will also contribute to increasing language diversity in high-quality data available to AI model developers.

As noted earlier, this reflects a broader paradigm shift. While the earlier focus of digitization initiatives was on enabling discovery and access to individual works, the development of LLMs has highlighted the importance of having large collections available in aggregate. The EBDC addresses this need by creating a shared mechanism for assembling, processing, and providing such collections at scale.

In summary, the proposed EBDC would create value for multiple stakeholder groups. Libraries gain access to a shared infrastructure under their control, reducing dependence on external providers and ensuring the provenance of digitized works. Researchers benefit from a reliable and centralized source of digitized public domain books. AI developers obtain high-quality, language-diverse training datasets with clear provenance. From a policy perspective, the EBDC aligns with the objectives of the EU data sovereignty objectives, strengthens the digital commons, and advances AI strategy goals.

High-level technical architecture

The core of the EBDC will be an online storage service that holds high-quality book scans (images), full-text files, and associated metadata. The storage system needs to handle large volumes of data (high-quality image files can exceed 1 MB per scanned page).

The storage service should be architected to allow distributed storage across facilities owned and operated by different entities. Access to the stored data will be available through a unified layer that exposes a set of APIs—regardless of where the stored data resides.

On the input side, a centrally defined data-processing pipeline (or set of pipelines) will ensure that all data in the EBDC meets minimum technical quality requirements. In the initial version, the pipeline must support ingesting books via the Google Books Retrieval Interface (GRIN) available to library partners.

Ideally, the elements described above would build on solutions already deployed by the Institutional Data Initiative (IDI). Given the overlap in functionality, this will likely apply to the input-side data-processing pipelines. By contrast, the storage component will likely need to be developed independently due to European data sovereignty concerns, although interoperability at the output/interface layer will be essential.

In general, the development of the technological components should be closely coordinated with the IDI. This is a realistic expectation given IDI's focus on making its technological building blocks available as open-source software to enable adoption by the global library community.

See Annex 1 for a more detailed list of elements of the EBDC that require a technical specification.

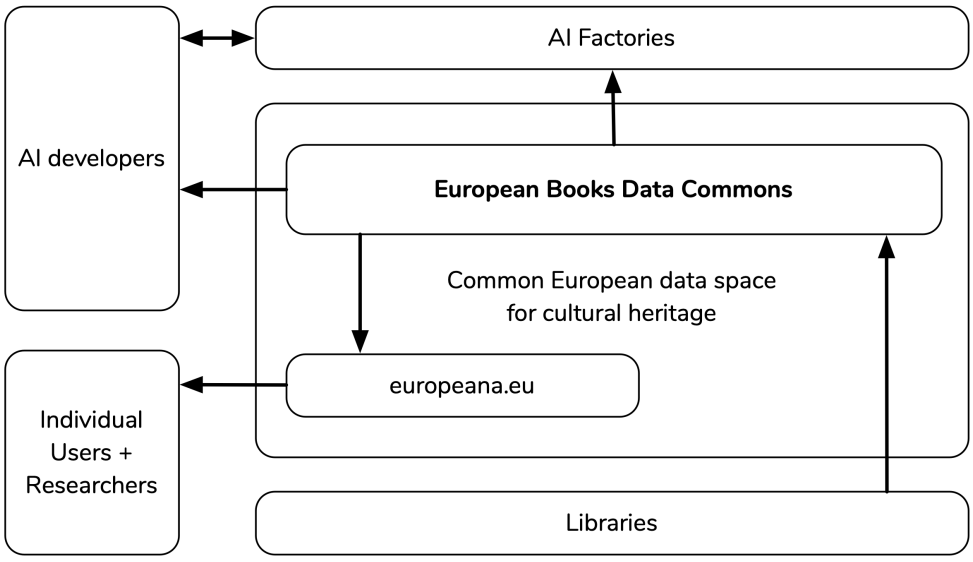
The EBDC in the European cultural heritage context

The European Books Data Commons will not stand on its own. Over the past two decades, Europe has invested substantially in digitizing cultural heritage and building an institutional arrangement to facilitate digital transformation in this sector. At the core of this arrangement are Europeana and the common European data space for cultural heritage (hereafter "the data space").

The EBDC should be seen as an additional element of this arrangement—an independent service within the data space alongside other components such as europeana.eu, which currently provides access to over 59 million digitized items shared by cultural heritage institutions across Europe. While Europeana primarily aggregates metadata related to a wide range of cultural heritage objects, the EBDC aggregates full digital artifacts—images, full text, and metadata—limited to books and similar written materials. We foresee technical integration between these data space services and compliance with overall governance. However, how deeply these services will be integrated into the data space stack remains an open question at this stage.

The European Commission's AI Continent Action Plan, published earlier this year, makes it clear that the data spaces—including the cultural heritage one—are increasingly seen as providers of high-quality data for the European AI ecosystem. This ecosystem centers on the so-called "AI Factories" and is supplemented by "Data Labs." This relationship is further elaborated in the Commission's Data Union Strategy. The EBDC, as presented here, can be understood as a concrete, sector-specific example of a Data Lab that supports the ambition to make Europe's cultural heritage a strategic asset for the European AI ecosystem.

The following diagram presents an idealized overview of how we envisage the European Books Data Commons fitting within this arrangement.



Libraries would provide collections that meet the EBDC's inclusion criteria: high-quality scans, full text, and free from copyright restrictions. The EBDC would process and store the data, then make it available to the public through several interfaces. Europeana.eu could serve as an interface that exposes the data to individual users, including researchers. The EBDC would also make bulk data available to AI Factories and directly (or via platforms such as Hugging Face) to other AI developers.

Governance + Sustainability

Positioning the EBDC as an independent entity that operates within the context of the cultural heritage data space raises questions about the governance of the EBDC itself. To fulfill its envisaged role, and in line with established principles of commons-based dataset governance, the EBDC must operate with a clear public-interest mission and be governed by the contributors to the commons—i.e., the contributing libraries. See Annex 1 for a more detailed list of elements that would need to be addressed by a governance framework.

In addition to institutional governance, the EBDC needs to provide a governance framework for access to the book datasets, based on conditional openness (or gated access). Such a framework should distinguish between the main categories of users and uses: research and other non-commercial purposes, open-source model development, and commercial model development. In the EBDC context, conditional openness means that datasets are freely accessible for non-commercial, public-interest uses, while commercial access may be subject to additional conditions— such as registration, purpose declaration, and payment where appropiate. Such contributions from commercial users could provide a means to support the long-term sustainability of the EBDC.

With regard to sustainability, it is clear that setting up and running the EBDC as outlined above will require substantial resources. This includes expenses for personnel (at least 4 FTE across software engineering and data-partner relations) as well as for running the infrastructure. A conservative estimate puts the annual operating costs between €500k and €750k.

Given that the high costs of running and maintaining infrastructure are one of the key reasons why some European libraries are not hosting their own Google Books collections, and given the lack of public funding instruments to ensure the long-term viability of public digital infrastructure, it will be essential to develop mechanisms that secure support from the envisaged user base (AI developers). This could take the form of direct contributions (sponsorship) or more structural arrangements (such as income from a levy on the use of publicly available information in commercially deployed AI systems, as we have suggested elsewhere). It is clear, however, that considerations about sustainability and how these can be ensured via a data-governance framework must be reflected in the overall setup of the EBDC from the start.

Annex 1: More detailed specifications

This annex provides a non-exhaustive overview of EBDC aspects discussed in this document that will require more detailed specifications.

Technical architecture

- Identifier and metadata standards: work/edition model, ISBN/ISCC, VIAF; formats like METS/ALTO, IIIF manifests, Dublin Core.
- Versioning and provenance: source, processing history, dataset releases, fixity (checksums), deduplication rules.
- Access and governance controls: authN/authZ if needed, rate limits, quotas, audit logs, abuse mitigation.
- Discovery: searchable catalog, OAI-PMH/IIIF discovery, bulk manifests, per-work and per-page endpoints.
- Delivery for AI use: bulk exports, shard formats, snapshots, content-addressed storage, streaming vs bulk download, CDN.
- Ingestion: provider submission interfaces, validation schemas, error reporting, normalization, derivative generation (OCR text, thumbnails, PDFs/EPUBs).
- Data quality: OCR pipeline, language/script detection, page order, layout metadata, QA metrics.
- Rights and policy signals: license/PD status, jurisdiction, TDM/opt-out flags, takedown workflow.
- Reliability: replication, mirroring across sites, backups, disaster recovery, lifecycle policies.
- Interoperability: cloud-based object storage, IIIF APIs for images, Europeana integration.

Governance

- Organizational form: consortium of libraries, foundation, or service operated by an existing body.
- Accountability: clear allocation of legal responsibility for compliance and risk management.
- Contributor role: governance rights tied to a provision of collections and resources.
- Transparency: public reporting on operations, finances, and data usage.

- Alignment: coordination with Europeana, the common European data space for cultural heritage, and related initiatives.
- Access framework: gated access model distinguishing between non-commercial/ public-interest use (free), open-source model development, and commercial model development.

Acknowledgements

This paper, authored by Paul Keller, builds on a series of structured conversations about the idea of a European Book Data Commons that were convened by the Europeana Foundation and Open Future during the first half of 2025. We thank the participants of these conversations for their time and input.

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture. The organization creates strategies for Digital Commons—democratically governed, collectively managed resources that provide an alternative to traditional ownership models. Open Future focuses on reimagining openness to foster a more balanced digital future that serves the public interest.

The Europeana Foundation is an independent, non-profit organisation that, as part of the Europeana Initiative, stewards the common European data space for cultural heritage and contributes to other digital initiatives that put cultural heritage to good use in the world. The Europeana Foundation promotes access to, and reuse of, cultural heritage and its work contributes to an open, knowledgeable and creative society.

Paul Keller is a co-founder and director of policy at Open Future. His work focuses on the intersection of copyright policy and emerging technologies. He works on policies and systems that improve access to knowledge and culture and protect the digital public sphere.