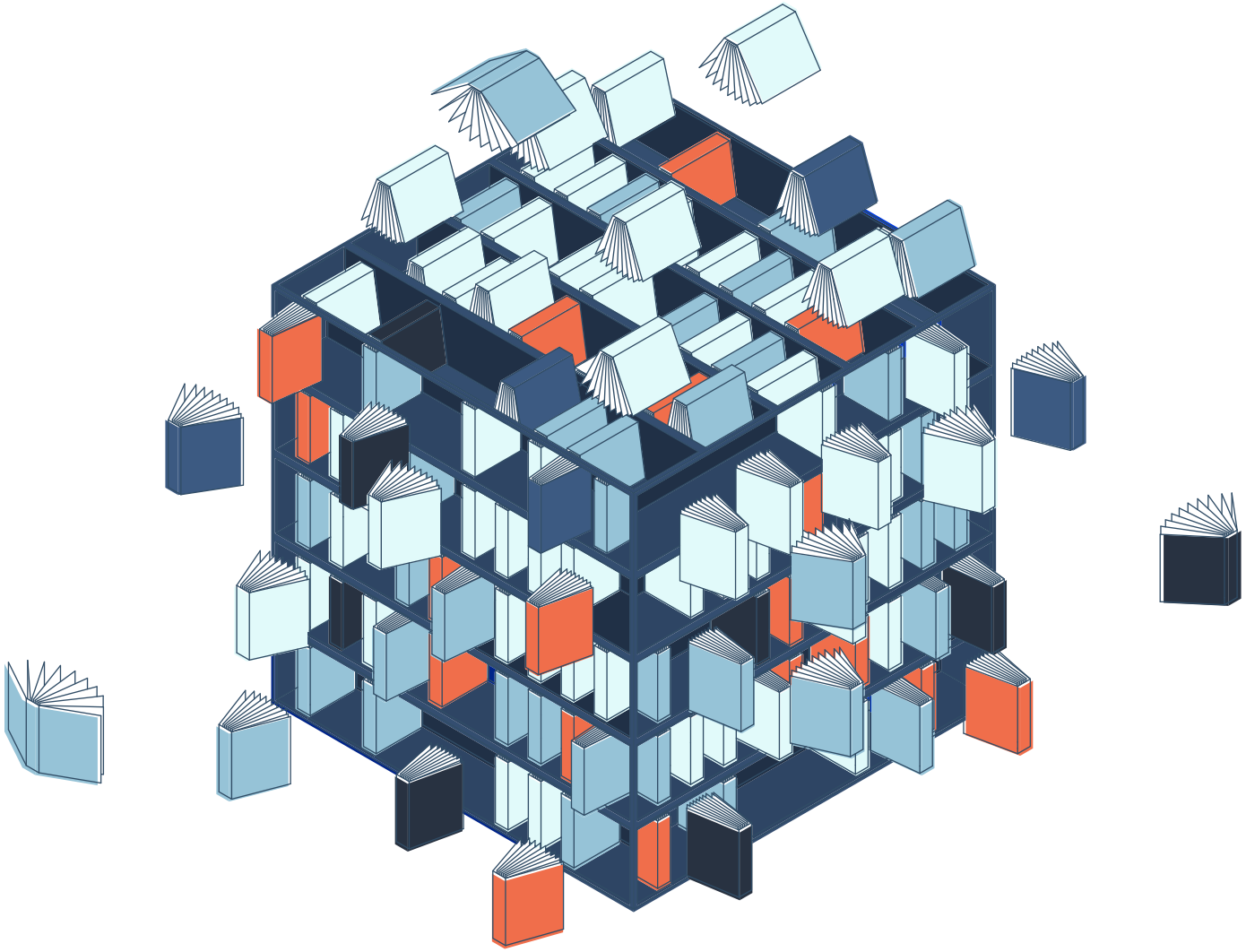
European Books Data Commons
feasibility study



KB nationale bibliotheek

europeana

July 2026

Foreword

The idea for a European Books Data Commons did not begin with this study. It grew out of earlier work developed with the European library community and first set out in a 2025 outline. The conversations we had with libraries during that phase made one thing clear: there is a growing need for shared infrastructure that lets Europe's libraries make their digitised collections available for AI development on their own terms, rather than leaving that to happen around them.

This feasibility study sets out how that infrastructure — the European Books Data Commons (EBDC) — can be brought to life. We commissioned it because we believe the EBDC offers a strong foundation for a library-led initiative, and because the moment to act is now. We come to it from complementary starting points.

For the KB, the initiative is about libraries keeping a hand in how AI systems represent knowledge:

> *"The European Books Data Commons initiative provides us with a unique opportunity to develop an AI-ready corpus under our own terms. As generative AI mediates more of what the public reads, sees and trusts, the question is no longer only whether libraries use these tools, but whether they have a hand in how the tools work: what they make visible, how they attribute and represent sources, and whose values they carry."* — **Wilma van Wezenbeek**, Director General, KB National Library of the Netherlands

For Europeana, it is the logical next step in decades of work on open access and interoperability:

> *"Europe's cultural institutions and Europeana have spent decades making collections accessible and data interoperable. As AI becomes the dominant interface to knowledge and culture, the next step is clear: make that data AI-ready within an open, shared corpus. The European Books Data Commons gives Europe a real opportunity to build an AI ecosystem grounded in public values, accountability and cultural diversity."* — **Harry Verwayen**, Director General, Europeana Foundation

The study that follows was prepared by Open Future at the request of the KB National Library of the Netherlands and the Europeana Foundation. It is based on desk research and interviews with fourteen experts and stakeholders from across the European library sector and the AI ecosystem. We hope that many libraries across Europe will join us in turning this proposal into a shared reality.

1 Introduction

This document identifies a pathway to establishing the core of the European Books Data Commons: a shared infrastructure that would make the full text of millions of digitised public domain books held by libraries across Europe available for re-use in forms optimised for AI developers and researchers working with large-scale language datasets. If implemented, the EBDC would constitute a significant contribution to the European Commission's 2025 Data Union Strategy, which aims to ensure that European AI developers have access to high-quality data including cultural heritage collections.

The remainder of the document is structured as follows. Section 2 sets out the proposition — the demand for digitised public domain books and the supply-side constraints that currently prevent access to existing collections. Section 3 presents the recommended implementation scenario, arrived at through a process of developing, testing, and progressively refining a set of options with library partners and the steering group. Section 4 addresses the governance of the system and its relationship to related European initiatives. Section 5 outlines the way forward, including a two-track implementation approach, an integrated timeline, and indicative cost estimates. Section 6 makes the case for acting now to build the European Books Data Commons.

2 The EBDC proposition

European national libraries collectively hold digitised collections of millions of public domain books in dozens of languages, covering centuries of cultural production. The greater part of this material exists in machine-readable form as a result of digitisation programmes carried out over the past two decades, in many cases in partnership with Google. It is largely inaccessible for bulk re-use: held in fragmented institutional silos, constrained by the terms of digitisation agreements, and not structured for the kind of large-scale computational access that AI development and digital humanities research now require.

The European Books Data Commons (EBDC) is a proposal to change this. Its core proposition is to give European AI developers and researchers access to something they currently cannot obtain elsewhere: a large-scale, multilingual, well-documented corpus of digitised European books, processed to consistent quality standards and made available under clear legal conditions. This would be achieved not by creating new digitisation programmes but by building a shared processing and distribution layer on top of collections that libraries already hold.

The proposition rests on two conditions: that there is genuine demand for this kind of data, and that the supply-side constraints that currently prevent access are addressable. The following section examines the demand side. The supply-side constraints — principally the embargo conditions in existing digitisation agreements — are addressed in section 2.2.

2.1 Demand for digitised public domain books

2.1.1 The AI training data market

Internal strategy documents from Anthropic, filed as part of US litigation in early 2026, provide an unusually direct window into how the AI training data market for books is currently structured[1]. Anthropic's internal assessment rates books as the highest-value category of written training data — more useful per token than web content, social media, or research databases — and confirms that book data measurably improves model performance. Because these are internal budget-allocation decisions rather than public-facing claims, they are a more credible signal of how much the company actually values book data.

The same documents describe a commercial acquisition pipeline that is structurally inefficient and running into supply constraints. Anthropic's primary book acquisition programme involves purchasing used physical books from wholesale vendors, scanning them, and extracting text. After several months of purchasing, the number of unique titles available per order was already declining. The document estimates approximately 130 million unique books exist worldwide, of which only a fraction are commercially purchasable. Anthropic explicitly identifies libraries as holding the majority of the remainder — material that is out of reach for commercial acquisition pipelines.

The commercial market has a further structural skew: approximately 90% of books purchased through this pipeline are English-language fiction, because these are the cheapest and easiest to source in bulk. Non-English material and specialised subjects require more effort and cost more to acquire. As the commercially available supply of English fiction is exhausted, demand is expected to shift toward exactly the kind of multilingual, subject-diverse collections that European national libraries hold.

2.1.2 The multilingual gap

The multilingual character of European library collections is the EBDC's principal differentiator from comparable initiatives. Existing large-scale book datasets — including those assembled through the HathiTrust partnership and Common Corpus — are multilingual in principle but heavily weighted towards English given their North American institutional base. European national libraries collectively hold digitised collections in French, German, Dutch, Spanish, Italian, Portuguese, Polish, and dozens of smaller European languages, covering material that the commercial market cannot replicate at scale.

Interviews conducted as part of this feasibility study confirm that demand for non-English training data currently exceeds available supply. AI developers building European language models and academic researchers working on multilingual NLP face a real data gap that is not addressable through the commercial channels that suffice for English-language projects.

[1] The document, titled "Data Acquisition Opportunity 2025" and dated October 2024, was filed as Exhibit 4 to Document 554 in *Bartz v. Anthropic*, Case 3:24-cv-05417-AMO (N.D. Cal.), on 21 January 2026. It is marked "Highly Confidential — Attorneys' Eyes Only" by Anthropic but is now part of the public court record.

Independent corroboration of this demand comes from Mistral AI's 2026 policy white paper addressed to EU policymakers, which includes a specific proposal to establish a centralised, multilingual repository of public domain works to provide high-quality training data for AI models[2]. Mistral identifies the absence of such an infrastructure as a concrete gap in the European AI data landscape. The proposal is broader in scope than the EBDC — it encompasses art, historical records, and scientific data alongside books — and advocates for unrestricted access rather than the conditional model examined in this report. It is nonetheless notable as an explicit demand signal from a major European AI developer, made in a public-facing and policy-directed register rather than inferred from internal strategy documents.

2.1.3 A note of caution

The demand case requires a degree of precision. Interviews with libraries with significant practical experience in AI data development surfaced a useful distinction that the EBDC's value proposition needs to hold onto. This proposition distinguishes between two separate claims: that training on older linguistic registers can reduce language model performance, which is a real and acknowledged constraint; and that the information and knowledge content of historical texts retains genuine interest and value for AI development. This distinction matters because the two claims have different implications for the EBDC's case.

The more defensible framing is therefore not that public domain books are the highest-value training data in absolute terms, but that they represent a significant and currently underserved category of supply — particularly for non-English languages where the gap between available and required data is larger, and where the commercial market has no comparable substitute. The linguistic diversity argument and the knowledge content argument are both stronger than the raw language model performance argument, and the EBDC's case rests primarily on the former two.

2.1.4 The cost argument

A further development that strengthens the supply-side case is the sharp decline in the cost of processing digitised collections into training-ready data. High-quality OCR text can now be produced from scan images at approximately €0.002 per page using current open-source models, putting the re-processing of a 100,000-page collection at around €200[3]. The technical barrier that once made this work the exclusive domain of well-resourced institutions or commercial platforms has effectively disappeared.

This changes the nature of the scarce asset. The OCR text layer is now reproducible at low cost by any party with access to the underlying scan images. The upstream asset — the scan corpus itself — remains controlled by libraries and in many cases has not been made publicly available. Libraries that have completed digitisation programmes hold the input from which training-ready data can be derived; the EBDC's value lies in making that derivation step

---

[2] Mistral AI, *European AI: A Playbook to Own It*, 2026, https://europe.mistral.ai/. The archive proposal is action #22, "Create a centralised and AI-ready archive for AI training and cultural preservation.

[3] Daniel van Strien, *Re-OCR'ing Collections*, 2026, https://danielvanstrien.xyz/posts/2026/re-ocr-collections/. The calculations underpinning our own cost estimates in section 5.3 are based on our own independent testing that builds on top of the work undertaken by van Strien.

systematic, consistent, and governed rather than in performing work that would otherwise be impossible.

2.2 Supply-side constraints on existing collections

European (national) libraries have been digitising their public domain holdings at scale since the mid-2000s. A large proportion of this work was undertaken in partnership with Google as part of the Google Books Library Project, under which Google scanned library collections in exchange for providing each participating library with a digital copy of the books it had contributed. Major European participants include the Bavarian State Library, the British Library, the Bodleian Libraries, the Koninklijke Bibliotheek, the Austrian National Library, and a number of Belgian and other institutions.

Across these partnerships, an estimated 4.5 million volumes[4] were digitised from European library collections. Significant non-Google digitisation has also taken place — most notably through the BnF's Gallica programme and Germany's retrospective bibliography projects — but the Google partnerships represent the largest single source of machine-readable public domain book data held by European libraries and are therefore the primary focus of the supply analysis in this section.

2.2.1 Agreement terms and access restrictions

The digitisation agreements between Google and its library partners are not uniform, but they share a common structural feature: they restrict the libraries' ability to make the digitised copies available for bulk re-use by third parties for a defined period after Google delivers the digital files. In most cases this restriction is framed as a temporary exclusivity arrangement. Available evidence from the KB–Google agreement and comparable arrangements suggests that the embargo period in most cases runs to 15 years from the date of delivery of the digital copy.

Article 12 of the 2019 Open Data Directive (ODD) limits exclusivity in digitisation partnerships to a maximum of 10 years (a period that can be extended based on a review of the underlying agreement). The 15-year embargo terms prevalent in the Google Books library agreements therefore exceed what the ODD permits — a discrepancy that has direct implications for how much of the digitised corpus is currently available for re-use, as set out in the following sections.

2.2.2 The current availability picture

The Google Books digitisation partnerships with European libraries were concluded at different points in time, with the bulk of scanning activity concentrated in the period from roughly 2008 to 2016. Given a 15-year embargo term, the first books digitised under these arrangements are only now beginning to exit the restriction period. At the time of writing, the number of volumes from European Google Books partnerships that are free from bulk re-use restrictions and could be made available through an infrastructure like the EBDC is relatively small — amounting to a fraction of the total digitised corpus.

[4] This and all subsequent figures in this section are based on collection data from interviewed library partners and publicly available figures for European Google Books Library Project participants.

2.2.3 The 10-year scenario

The picture changes materially if the restriction period were aligned with the 10-year maximum permitted under the Open Data Directive[5]. Under a 10-year term, a significantly larger proportion of the digitised corpus would already be free for re-use, because the restriction period on books scanned in the earlier years of each partnership would have expired.

Based on an analysis of the available collection data for the principal European Google Books partner libraries, we estimate that capping embargo periods at 10 years would make available for the EBDC a corpus of at least 2 million volumes that are currently subject to restrictions on bulk re-use. This figure covers the subset of Google-digitised volumes whose 10-year term would already have elapsed and which libraries could therefore contribute without renegotiating their agreements with Google. It does not include volumes currently free from restrictions, which would add further supply (estimated to be in the range of 0.5 to 1.5 Million volumes), nor does it include non-Google digitised collections, which represent a separate and additional source (estimated to be in the range of 1-2 million volumes).

Given that the Google-digitised corpus represents the largest single body of machine-readable public domain book data held by European libraries, a reduction of the embargo period to the ODD-compliant 10-year term is a precondition for the EBDC operating at meaningful scale in the near term. Based on ongoing conversations between Google and its partner libraries, it seems reasonable to expect that Google will offer new agreements to all partner libraries that would — with retroactive effect from the date of digitisation — reduce the embargo period to 10 years after receipt of the digital copy. This would bring a substantial proportion of the digitised corpus within reach for the EBDC in the near term, as reflected in the volume estimates above.

It would also remove the principal remaining barrier to libraries making these works available to large-scale re-users — and with that barrier gone, the case for doing so becomes difficult to resist. The final section of this report returns to that argument.

3 Implementation scenarios

This section presents the implementation scenario recommended by this feasibility study (section 3.3). It was arrived at through a process that began with a broad set of options and progressively narrowed the focus: first by eliminating one of two initial scenarios, then by expanding the preferred scenario in response to stakeholder feedback, and finally by paring it back in response to concerns about complexity and loss of focus.

The process started with two distinct implementation scenarios — a coordinated access model and a shared infrastructure model — which were developed and validated internally with the KB and Europeana. This validation produced a clear preference for the shared infrastructure scenario, for reasons set out in section 3.1.

---

[5] For a more detailed analysis of the relationship between the existing Google Books agreements and the Open Data Directive see: Arends, A. and Bont, A. de and Rosenberg, M (2024): Demonopolizing the European Public Domain: Google Books Exclusivity Clauses and the Open Data Directive.

Both scenarios were subsequently presented to representatives of eight (national) libraries and a number of stakeholders from the AI ecosystem, in order to gather feedback on the preferred approach. That consultation identified a range of additional requirements and access modalities that potential contributing institutions would want the shared infrastructure to support. The resulting expanded scenario, and the reasons it was ultimately set aside, are described in section 3.2.

The expanded scenario was assessed by the steering group, which raised concerns about the complexity of the resulting architecture and the risk of losing focus on the core value proposition. This led to the identification of a pared-down, phased implementation scenario — one that begins with the minimum viable realisation of the EBDC concept and defers more complex functionality to later, optional phases. That scenario is presented in section 3.3.

3.1 Coordinated access versus shared infrastructure

The initial analytical step was to define the range of possible implementation approaches. These reduce to two fundamentally different models. In a coordinated access model, the only shared element is a discovery layer: a unified catalogue or registry that aggregates metadata from participating libraries and allows users to locate digitised book collections across institutions. Processing, storage, full text, and access interfaces all remain with the individual participating libraries. In a shared infrastructure model, libraries contribute their digitised collections to a jointly operated system that handles ingestion, processing, storage, and access, and delivers the output as structured, cross-institutional datasets.

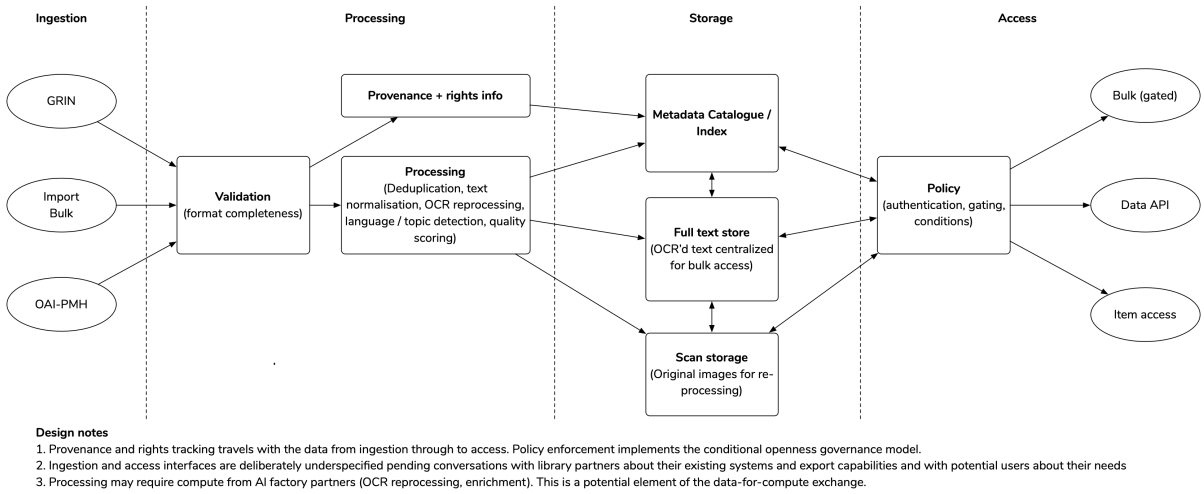
The coordinated access model has the obvious advantage of requiring minimal commitment: no new infrastructure, no transfer of data outside institutional control, and low operational costs. On closer examination, however, it does not deliver on the core value proposition of the EBDC. Users seeking large-scale, consistently processed, multilingual book corpora would still face the same fragmented landscape that exists today — navigating different APIs, incompatible formats, varying quality levels, and inconsistent access conditions across individual library systems, assuming those systems exist at all. For many European libraries, functional bulk access endpoints do not currently exist. A coordinated access EBDC would also offer no meaningful added value over what Europeana and the common European data space for cultural heritage (from now on ‘the data space’) already provide through its metadata aggregation services. It is not a viable path for the EBDC as a distinct initiative.

The shared infrastructure model is the only approach that delivers the functionality the EBDC concept was designed around: a processing pipeline that turns heterogeneous digitised collections from across Europe into consistent, usable, well-documented corpora, stored on jointly governed infrastructure and made accessible under clear conditions. It requires more from participating libraries — a commitment to contribute collections and to operate within a shared governance framework — but it is the only model that solves the underlying problem rather than cataloguing it. The remainder of this section therefore proceeds on the basis of the shared infrastructure approach.

3.2 Shared infrastructure — diverging demands

The shared infrastructure model that emerged from the initial validation with the KB and Europeana is built around three functional stages. A processing pipeline ingests digitised collections from contributing libraries — primarily via the GRIN export interface used in the Google Books partnerships — and produces structured, quality-controlled datasets as its output. A storage stage holds the processed data: metadata, full text, and scan images. An access stage delivers the EBDC's two data products: pre-packaged datasets available for bulk download, and individual item access — allowing patrons and researchers to retrieve individual books with metadata, full text, and scans — through an API, a dedicated interface, or via integration with existing platforms such as Europeana and the data space. These three stages together constitute the EBDC as a distinct service — one that takes fragmented digitised collections as inputs and delivers a consistent, well-documented cross-institutional corpus as output.

This model was developed in collaboration with Europeana and the KB and then presented to the broader group of interviewed libraries and stakeholders to gather feedback on its design and assess the conditions under which they would be willing to contribute.



3.2.1 Diverging demands from the stakeholder interviews

The stakeholder interviews validated the core logic of the shared infrastructure model but surfaced a wide range of additional requirements and access modalities that potential contributing institutions would want the infrastructure to support. These requirements do not point in a single direction — in some cases they are in direct tension with each other.

On access governance, the consultation revealed a fundamental division. Several institutions expressed strong preferences for conditional access arrangements: the ability to attach conditions to the release of specific datasets, whether for reasons of revenue generation, control over commercial use, or the ability to handle in-copyright material alongside public domain collections. Others held unrestricted openness as a near-absolute requirement for participation, on the grounds that public domain material should not be subject to access

restrictions regardless of who is requesting it or for what purpose. These positions are not easily reconciled within a single service design.

The interviews also revealed varying levels of interest in the individual item access component. For several libraries, the ability to offer their patrons and researchers access to processed individual books through the EBDC was an important part of the value proposition — both as a service to their own user communities and as a demonstration that the EBDC delivers benefits beyond the AI developer use case. For others, individual item access was a secondary concern: what justified contributing collections was the bulk dataset output, and per-item retrieval was at best a useful by-product. This divergence has direct implications for how much investment the individual item access stage warrants at the outset.

A structurally distinct requirement came from institutions that already operate their own processing infrastructure. Rather than depositing raw scan material for the EBDC to process, these institutions indicated a preference for delivering already-processed, standards-compliant datasets directly to the EBDC's storage and access stage — effectively treating the EBDC as a delivery endpoint rather than a processing service. This model, in which the processing pipeline can be bypassed by institutions with sufficient technical capacity, would allow libraries with existing pipelines to participate on their own terms while still contributing to the shared corpus.

3.2.2 The resulting complexity

Accommodating these requirements in a single architecture would produce a substantially more complex system than the initial shared infrastructure model. The bypass pipeline pathway — allowing institutions with their own processing infrastructure to deliver pre-processed data directly to the EBDC's storage and access stage — is technically feasible but introduces significant governance questions about quality standards, documentation requirements, and accountability.

The more fundamental complexity, however, comes from the conditional access dimension. The interviews revealed that libraries have widely diverging ideas about what conditions they would want to impose on dataset access: ranging from registration and purpose declaration, to restrictions on commercial use, to revenue generation, to guarantees that data will not be shared further downstream. These conditions are not equivalent in what they require technically. Some — such as commercial use restrictions and purpose declarations — can be handled through contractual access agreements. Others — in particular revenue guarantees and downstream sharing controls — cannot be reliably enforced through limitations or conditions imposed on access and would require significantly more complex technical implementations, such as on-premises data access and processing environments. These are costly to build and complex to maintain, and would in practice serve only a small number of contributing institutions with specific requirements. Building them into the core architecture would risk fragmenting the EBDC's offering and undermining the consistency and openness that constitute its core value proposition.

This expanded scenario was assessed by the steering group, which raised concerns about complexity and the risk of losing focus on the core value proposition. These concerns are

addressed in section 3.3, which presents the pared-down phased implementation that this feasibility study recommends.

3.3 The path forward: a phased approach

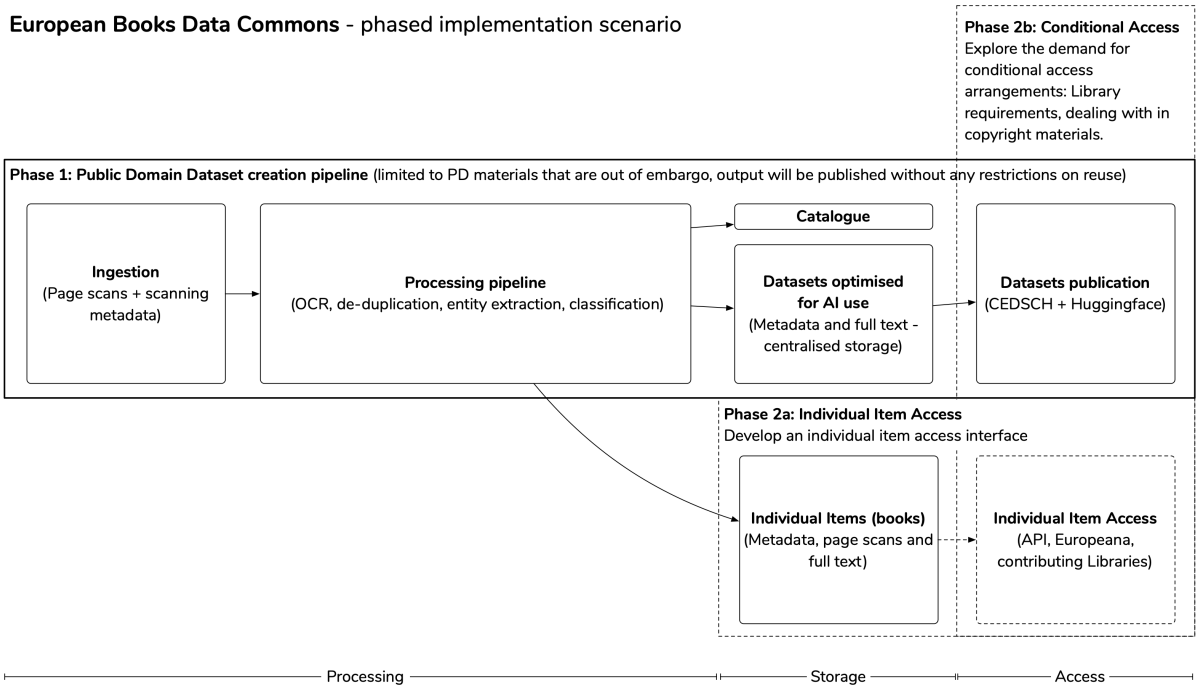
The stakeholder interviews surfaced a wide range of positions on what the EBDC should be and under what conditions libraries would be willing to contribute — spanning both the access conditions that contributing institutions would want to impose and the modalities through which different user communities would want to retrieve data. Accommodating that full range would produce a system too complex to build coherently and too constrained to deliver on its core purpose. The phased implementation described below is designed around that core purpose.

In addition there is also a structural reason to be cautious about building features in response to articulated preferences. The experience of existing European aggregation projects — most notably Europeana and the data space— suggests that an approach which develops functionality based on what potential contributors say they need does not reliably work in practice. Institutional priorities shift, governance arrangements change, and functionality built to meet specific conditions frequently goes unused once the infrastructure is live. Building for a wide range of expressed needs at the outset is therefore not only complex but also likely to result in an infrastructure that is over-specified for the actual use that materialises.

Given these considerations, a technical architecture designed to accommodate the full range of stakeholder positions is not considered feasible as a starting point. The recommended approach is instead a phased implementation that begins with the core value proposition and defers more complex functionality to later phases, each of which can be realised only after the core infrastructure has been successfully deployed and has demonstrated the underlying case for the EBDC.

The three phases are structured as follows. In the first phase, the EBDC consists of a processing pipeline that ingests digitised public domain books contributed without any restrictions on re-use, processes them into structured datasets, and publishes those datasets without access restrictions. This phase is the minimum viable realisation of the EBDC concept and the precondition for everything that follows. In a second, optional phase, the EBDC can be expanded by an individual item access stage that provides access to the digitised books in the collection — metadata, full text, and page scans — through an API, a dedicated interface, or both. In a third, optional phase, a conditional access infrastructure can be built that allows the release of specific datasets under defined conditions. This phase would only be implemented if there is substantial and credible demand from data providers who can commit to contributing significant collections, but would only do so if conditional access functionality exists.

The diagram below illustrates the three phases and the open questions that each raises. These questions are addressed in turn in sections 3.3.2 through 3.3.4.



3.3.1 Data products and dataset releases

The primary output of the EBDC is pre-packaged datasets: curated, documented, versioned corpora that users can download as defined units. Datasets are composed along two primary axes — language and time period — with secondary dimensions including OCR quality level and subject classification where derivable from metadata. Each dataset release is a versioned snapshot with a persistent identifier, enabling reproducibility in research and provenance tracking in AI training pipelines.

The EBDC is not a one-time digitisation effort but a continuously operating service. As new collections are contributed and ingested, and as OCR technology improves, new dataset versions are released on a regular cycle. This continuous release model means the corpus grows in volume and improves in quality over time, and that users can track changes between versions. It also means that the processing pipeline is a permanent operational function of the EBDC, not a project-based activity with a defined end date.

3.3.2 Core processing service

The core processing service is the first phase of the EBDC and the foundation on which everything else is built. It consists of three sequential stages: ingestion, processing, and dataset publication.

In the ingestion stage, contributing libraries provide page scans of digitised books together with associated scanning metadata. The primary input pathway is via the GRIN export

interface available to Google Books partner libraries, which provides a standardised mechanism for extracting this material. The pipeline will also need to support direct bulk upload and other delivery mechanisms for libraries that digitised their collections independently of Google.

In the processing stage, the scanned page images and associated metadata are passed through a pipeline that performs OCR to produce structured full text, followed by de-duplication at the volume level, entity extraction, language detection, and classification. The output is structured full text with associated metadata and quality annotations. Processing documentation is generated as a formal output alongside the data, recording which models and parameters were applied — both to enable downstream users to understand what they are working with, and to ensure that future re-processing as OCR technology continues to improve can build on rather than repeat prior decisions.

The processed output — datasets of metadata and full text optimised for AI use — is then published without any access restrictions. The primary publication channel is EBDC-controlled infrastructure, with HuggingFace providing a secondary interface for the AI and ML research community. In phase 1, datasets are limited to public domain material and no conditions are attached to re-use.

Open design questions

- **Ingestion standards:** Beyond GRIN output, what technical standards or requirements should contributing libraries meet when delivering scan material? This needs to be defined in conversation with potential founding partners before the pipeline is built.
- **Output format:** The default output format for bulk datasets should be JSONL or Parquet, both standard in AI development workflows. The choice between them, and any additional formats to support digital humanities use cases, requires agreement with the intended user communities before the pipeline specification is finalised.

3.3.3 Individual item access

The second phase would extend the EBDC by adding an individual item access interface alongside the bulk dataset output of phase 1. This interface would allow users — including the contributing libraries and their patrons, researchers, and other users — to retrieve all resources associated with a specific book: its metadata, full text, and page scans. This is distinct from bulk dataset access in that it serves per-work queries rather than large-scale download, and requires a different technical interface.

The individual item access interface could be delivered through one or more channels: a dedicated EBDC API, a user-facing EBDC interface, or integration with existing platforms such as Europeana. The most likely initial implementation would be an API, which both enables integration with Europeana and allows contributing libraries to surface EBDC-processed versions of their own collections through their own platforms. Phase 2a draws on the individual item data — metadata, scans, and full text — stored by the phase 1 production pipeline from the point of ingestion.

Open design questions

- **Interface design:** Should phase 2a prioritise an API for institutional integration, a public-facing EBDC interface, or both? The answer depends on which user communities are being served and what Europeana's integration capacity looks like at the time of implementation.

- **Europeana integration:** Europeana already provides an interface for individual item access and is the natural integration point for surfacing EBDC content to end users. The extent to which the EBDC should build its own interface versus relying on Europeana needs to be agreed before phase 2a is initiated.

3.3.4 Conditional access

Phase 2b would explore the conditions under which specific datasets could be made available under defined access conditions, rather than the fully open publication model of phase 1. This phase would extend the EBDC to support use cases that phase 1 cannot accommodate: libraries that are willing to contribute collections only under specific conditions, and potential future integration of in-copyright or out-of-commerce material alongside the public domain corpus.

The conditions that could be supported through a standard access infrastructure — registration, purpose declaration, restrictions on commercial re-use — are technically straightforward and could be implemented through contractual data access agreements rather than complex technical controls. Where technical enforcement is required, a lightweight EBDC-controlled access gate, implemented on top of the existing EBDC S3 infrastructure, provides the appropriate mechanism. This is the appropriate scope for a phase 2b conditional access infrastructure. As established in section 3.2, conditions that require stronger technical enforcement — such as revenue guarantees or downstream sharing controls — fall outside what this phase is designed to deliver.

Phase 2b would only be initiated if there is substantial and credible demand from data providers who can commit to contributing significant collections contingent on this functionality. It is explicitly conditional: if that demand does not materialise, phases 1 and 2a constitute the complete EBDC as deployed.

Open design questions

- **Feasible conditions:** What specific access conditions can a phase 2b infrastructure credibly support, and which are out of scope? A clear statement of what the EBDC's conditional access infrastructure will and will not do is a prerequisite for engaging with libraries that have expressed interest in this model.

- **Governance of conditions:** Who decides what conditions are permissible, and how are disputes between contributors and users resolved? The governance framework for phase 2b is more complex than for phases 1 and 2a and needs to be designed before implementation begins.

- **In-copyright and out-of-commerce material:** Could phase 2b provide the infrastructure for making out-of-commerce works available under the DSM Directive (Articles 8–11)

framework alongside public domain material? This would substantially expand the temporal scope of the corpus but introduces a different rights management regime.

The implementation scenario described in this section reflects a deliberate narrowing of focus. The stakeholder interviews confirmed that there is willingness among European libraries to contribute to a shared infrastructure, but also that the conditions attached to that willingness are widely divergent and cannot all be accommodated in a first phase without compromising the core proposition.

The recommended starting point is therefore an infrastructure with a clearly bounded scope: digitised public domain books go in, openly available datasets come out. The goal is to ensure that AI developers and researchers working in Europe have access to as much of Europe's digitised public domain book heritage as possible — processed to consistent quality standards, documented, versioned, and available without restrictions on re-use. Everything else — individual item access, conditional access, in-copyright material — can be added once that foundation is in place and has demonstrated its value. Section 5 below sets out the practical steps required to get there, including a two-track approach that combines a small-scale pilot with the build-out of the full production system.

4 Governance of the EBDC

The infrastructure described in section 3 should be owned and controlled by the libraries that contribute to it. This is not only a matter of institutional preference — it is a functional requirement. Libraries have a direct interest in ensuring the accuracy and provenance of the digitised collections they hold, and in being able to set the terms under which those collections are made available. Even in phase 1, where datasets are published without access restrictions, the conditions attached to contribution — what quality standards apply, how provenance is documented, how the corpus is governed over time — are decisions that should rest with the contributing institutions rather than with a third-party operator. A library-led governance model is also likely the only arrangement that gives contributing institutions the confidence to transfer collections to shared infrastructure and commit to maintaining that contribution over time. Without it, the EBDC risks replicating the structural dependency on external operators that it was designed to remedy.

Two principles follow from this that apply to any involvement of commercial entities in the EBDC, whether as infrastructure providers, processing partners, or distribution channels. First, collaborations with commercial entities must not result in (new) forms of exclusivity over EBDC data. Second, they must not result in (new) infrastructural dependencies on commercial providers. These red lines apply regardless of the form the collaboration takes — they are not negotiating positions but boundary conditions for the governance model.

It is useful to distinguish two functions that need not rest with the same body. The first is governance: ultimate authority over the EBDC's direction, the terms on which collections are contributed and published, and the decision on the EBDC's legal and operational form. Consistent with the principle set out above, this authority rests with the contributing libraries, acting through a governance structure they establish. The second is the operator function: running the infrastructure and providing the day-to-day coordination the service requires. The

Europeana Foundation is well positioned for the operator role, given its pan-European mandate, its role as operator of the data space within which the EBDC sits, and its existing relationships across the European library community; it has indicated it would be a candidate. The KB, as a co-initiator of this study and a prospective founding partner, has a central role in establishing the governance structure but is not a candidate to host the infrastructure: a national library is not the appropriate base for European-level infrastructure of this kind.

The operator decision is subordinate to, and made by, the library-led governance structure. A distinct question is where legal personality sits — whether the governance structure is itself constituted as the legal entity that contracts with contributing libraries, or whether contracting is delegated to the operator. This should be settled deliberately rather than defaulting to the operator, since the entity that holds the contracts with contributing libraries and has custody of the corpus accrues practical control over the EBDC irrespective of the governance principle. The available forms — a new legal entity, a hosted structure within an existing institution, or a distinct component within the common European data space for cultural heritage — remain to be agreed and are a prerequisite for the timeline in section 5.

4.1 Relationship to other initiatives

The EBDC does not operate in a vacuum. A range of initiatives at European and national level pursue overlapping objectives or work in the same space, and the EBDC's relationship to them is not uniform: in some it operates within an existing framework and builds directly on it; with others it occupies a distinct position that needs demarcating to avoid duplication; and others again serve as reference points for its own design. The following sections set out how the EBDC should relate to each.

The EBDC's relationship to these initiatives follows from the intended positioning of the EBDC as a self-contained initiative — governing, funding, and operating itself — that sits within the common European data space for cultural heritage rather than alongside it. This setup is deliberate. Self-governance and an independent funding base are what give contributing libraries control over the corpus and what keep the EBDC clear of the dependencies set out at the start of this section. Membership of the data-space ecosystem is what gives the EBDC access to resources it could not credibly assemble on its own: processing compute from the AI factories, the institutional standing that comes with being part of the data space, and the technical and relational infrastructure that Europeana already operates. This position inside the ecosystem, rather than as a standalone initiative, makes the case for AI-factory compute for the one-time re-OCR of the corpus a realistic ask within the EU AI ecosystem.

4.1.1 The common European data space for cultural heritage

The EBDC is designed to operate within the common European data space for cultural heritage (from now on ‘the data space’). The data space provides the infrastructure through which cultural heritage institutions make large scale datasets available for re-use; the EBDC builds on that infrastructure by adding a processing and distribution layer specifically optimised for large-scale computational re-use of digitised book collections. In this sense the EBDC and the data space are complementary by design: the data space provides the

governance framework and institutional embedding into the EU data ecosystem, while the EBDC provides the specialised pipeline that turns digitised collections into AI-ready datasets and channels them toward the AI ecosystem. As operator of the data space, the Europeana Foundation builds and maintains relationships at pan-European level — including with the AI factories and data labs — that the EBDC can draw on to reach the compute and distribution partners its model depends on.

4.1.2 The EU AI ecosystem (Data Labs and AI Factories)

The European Commission's AI infrastructure programme envisages a network of data labs whose function is to connect the data spaces with AI factories: processing and preparing datasets drawn from the data space into forms suitable for use in AI development, and channelling those datasets toward the compute infrastructure where models are trained. This is precisely the function the EBDC is designed to perform for European book data — ingesting digitised collections from contributing libraries, processing them into AI-ready datasets, and making them available without restrictions on re-use.

The EBDC should therefore position itself explicitly as a proto-data lab: an initiative that already operationalises the data lab concept for a specific and significant category of cultural heritage data, ahead of the formal data lab network being established. This positioning has two practical implications:

- **First**, AI factories operating under the European AI infrastructure programme are the natural source of the processing compute required to run the EBDC pipeline — the data-for-compute exchange model, in which AI infrastructure partners provide processing capacity in return for access to the resulting datasets, is the preferred arrangement and can significantly reduce processing costs.

- **Second**, as the data lab network takes shape, the EBDC should seek recognition as part of the data lab network, with data labs providing a level of stable infrastructure and staff funding over the longer term. The EBDC engages this process directly, as a proto-data lab arguing for its own recognition as the network's architecture and mandate are fixed. Its integration with the data space (section 4.1.1) adds a second, complementary channel: as operator of the data space, the Europeana Foundation is well positioned to engage the Pharos-led Language and Culture Data Lab on behalf of the cultural heritage sector as a whole, strengthening the EBDC's institutional pathway into the network without it depending on that channel alone.

Engaging with the data lab design process at this stage, before the architecture and mandate are fixed, is therefore a near-term priority.

4.1.3 European Research Infrastructures (EOSC / CLARIN / DARIAH)

The EBDC should be positioned as an initiative that operates within the data space logic and does not depend on making commitments to existing European Research Infrastructures such as CLARIN or DARIAH. The EBDC's primary relationships are with the data space, the data lab network, and the library community — not with research infrastructure organisations whose mandates and governance structures are oriented toward different objectives. This

positioning avoids duplication and keeps the EBDC's governance commitments focused on the institutions most directly relevant to its mission.

4.1.4 National initiatives

There is likely to be some overlap with national-level initiatives. Having set aside the idea of a bypass pipeline for pre-processed data, incorporating material from existing national initiatives into the EBDC requires contributing libraries to provide book scans to the EBDC processing pipeline. While defining the terms of these relationships is not a primary focus of the initial phase, the EBDC will need to develop clear arrangements for how national initiatives can contribute — covering quality standards, provenance documentation, and governance participation — before such partnerships can be formalised. Given that several potential contributing institutions already operate significant national digitisation programmes, this question is likely to arise early and should be treated as a near-term design item rather than a deferred one.

4.1.5 Non-European initiatives

Two non-European initiatives serve as reference points for the EBDC, each relevant in a different respect. HathiTrust represents the closest analogue on the library aggregation and access side: a jointly governed repository that aggregates digitised collections from major research libraries, a significant proportion of which were digitised through the Google Books Library Project. The Institutional Data Initiative (IDI) at Harvard represents the closest analogue on the data processing and AI re-use side: a programme that has demonstrated the technical feasibility of extracting and publishing library book data at scale for AI development purposes.

HathiTrust and the EBDC share the broad principle of library-controlled infrastructure as an alternative to commercial operator dependency, and for US libraries that participated in Google Books, HathiTrust serves broadly the function the EBDC is designed to serve for European libraries. The operational contexts are, however, substantially different. HathiTrust's remit is shaped primarily by the constraints of the Google Books litigation settlement, which governs how it can handle both in-copyright and public domain material and significantly limits its ability to make collections available for bulk re-use. It is not a model the EBDC should seek to replicate, and significant operational overlap is not expected.

IDI is a more directly relevant reference point for the EBDC's near-term technical work. Its development of the grin-transfer toolchain — already validated on over one million volumes from Harvard Library — provides the primary ingestion mechanism for the EBDC pipeline, and its published dataset (Institutional Books 1.0) demonstrates that GRIN-based extraction and publication of AI-ready book corpora is achievable at meaningful scale. Collaboration with IDI on tooling and standards development is anticipated as the EBDC moves from concept to implementation.

A third non-European reference point is the Public Interest Corpus initiative, a US-based effort led by Authors Alliance and Northeastern University with Mellon Foundation support, which published an implementation framework in early 2026. The initiative shares the EBDC's core premise — that research libraries are uniquely positioned to provide large-scale, high-quality

book data for AI development and computational research, and that the current concentration of access to such data in a small number of commercial actors is a problem requiring a library-led response. Its conclusions about demand, the value of multilingual and historically diverse collections, and the case for library-governed infrastructure broadly validate the assumptions underlying the EBDC.

However the two initiatives operate in substantially different legal and institutional contexts: The Public Interest Corpus is designed primarily around the US fair use framework and encompasses both public domain and in-copyright material, with access mediated through institutional agreements with academic and non-profit users. The EBDC's Phase 1 proposed here is limited to public domain books which both simplifies the legal position and removes the need for the access restriction infrastructure that the Public Interest Corpus requires. The PIC implementation framework is also relatively high-level on technical and operational specifics. Given the complementary rather than overlapping scope, coordination on standards, tooling, and the broader case for library-led infrastructure is worth pursuing as both initiatives move toward implementation.

4.2 Contribution terms and library entitlements

Contributing libraries are the foundation of the EBDC. Without their willingness to provide collections and commit to shared governance, the infrastructure cannot exist. The terms on which libraries contribute therefore need to be clear from the outset — both to give contributing institutions confidence in what they are committing to, and to ensure that the governance model is sustainable over time.

The core contribution term for Phase 1 is straightforward: libraries provide digitised collections (scanned books) without restrictions on the re-use of the resulting datasets. This is a precondition of the EBDC's core value proposition — openly available datasets cannot be produced from collections that carry re-use conditions — and it is the basis on which the infrastructure is designed. Libraries that are unwilling or unable to contribute on these terms are not excluded from the EBDC, but their collections cannot be incorporated into Phase 1 — they would need to await the conditional access mechanisms envisaged in Phase 2b.

In return for contributing collections on these terms, libraries are entitled to participate in the governance of the EBDC and to access the processed outputs of their own collections. Governance participation is addressed through the library-led governance structure described above. Access to processed outputs is a practical entitlement that needs to be established in the contribution agreement from the outset, even though the mechanism for delivering it is primarily a Phase 2a design question. The design decision to store individual item data from the point of ingestion in the Phase 1 production pipeline is partly motivated by this entitlement: libraries should not have to wait for a full re-processing run to access the outputs derived from their own collections. Access to the outputs derived from other contributors' collections is not a separate entitlement of this kind: in Phase 1 those outputs are published as openly available datasets, and in Phase 2a they are reachable through the same individual item access interface — available to contributing libraries on the same basis as to any other user.

5 Way forward

This section outlines the steps required to move the EBDC from concept to operational infrastructure. It proceeds in three parts: first, a description of the two-track implementation approach — a pilot followed by the full production system; second, an integrated timeline covering both tracks; and third, indicative cost estimates for both the pilot and the production system.

5.1 A two-track implementation approach

The EBDC will be built in two stages: a pilot implementation at small scale, followed by the full production system.

The pilot is designed to validate the core technical and organisational assumptions before significant resources are committed. It covers a collection of approximately 50,000 books drawn from one or more founding library partners, and implements the complete Phase 1 pipeline in miniature: ingestion via GRIN, processing using both the Google OCR baseline and a VLM-based re-OCR model, and publication of the resulting datasets without access restrictions. Running both processing approaches in parallel serves two purposes: it generates a direct quality comparison on actual EBDC collections rather than relying on generic benchmarks, and it validates the compute procurement assumptions that underpin the production cost model. For the latter purpose, the pilot will benchmark VLM re-OCR performance and cost across the two procurement scenarios available to a multi-library public institution: AI factory capacity (where available under terms consistent with the EBDC's governance principles) and commercial European cloud providers. Within each scenario, benchmarking will cover a range of GPU performance tiers, since the degree to which OCR workloads benefit from higher-end hardware has direct implications for production cost estimates. The results will constitute a standalone pilot output — a reproducible evaluation dataset that contributing libraries can refer to when making their own compute procurement decisions.

From the outset, the canonical data store is EBDC-controlled S3-compatible object storage. The pilot will also establish two discovery interfaces: a primary EBDC catalogue viewer built on a zero-backend architecture (DuckDB running in the browser, hosted as a static page on EBDC infrastructure), and HuggingFace as an additional interface for the ML community. The precise relationship between these two interfaces — and in particular whether HuggingFace should function as a viewer pointing to data stored on EBDC-controlled S3, or as a parallel storage and distribution channel in its own right — is identified as an open design question below.

The production system extends the pilot along two dimensions: scale (from 50,000 to up to 3 million volumes, ingested across multiple library partners) and completeness (individual item storage for Phase 2a, and the full enrichment pipeline including deduplication, language detection, and classification). The production system also stores page scans from the point of ingestion — a decision driven by the practical constraint that re-downloading from GRIN would require another multi-month extraction cycle. Conditional access infrastructure, if

introduced at all, would be implemented as a lightweight EBDC-controlled access gate and is deferred to Phase 2b.

Open design question:

- **Data storage and distribution architecture:** Two options are available for the relationship between EBDC-controlled S3 storage and HuggingFace. In the first option, S3 is the sole canonical data store and HuggingFace functions as a discovery interface pointing to it — preserving full infrastructure independence but potentially limiting discoverability and reuse in the AI research community, where HuggingFace is the dominant distribution platform. In the second option, data is stored on both S3 and HuggingFace, with S3 as the primary canonical store and HuggingFace as a genuine distribution channel — maximising reach but establishing a structural dependency on a commercial platform that may be difficult to unwind later. The choice between these options has implications for the EBDC's governance, identity as well as its operational model and should be resolved by the steering group before the pilot begins.

5.2 Timeline

Several factors make the timing for building the EBDC particularly urgent. Libraries are already dealing with unmanageable automated scraping of their online collections. AI developers need non-English language data that the commercial market cannot supply. The European Commission is actively pushing for cultural heritage data to be made available for AI development as part of its Data Union Strategy. And the data labs and AI factories that could serve as compute and distribution partners for the EBDC are being designed right now — the window to shape their architecture and mandate to include the EBDC is open, but will not remain so indefinitely.

Against this backdrop, a cautious, prolonged build-up process carries real risk. The value of the EBDC depends in part on being early enough to shape how European AI developers access book data, and early enough to be a credible partner in the data labs design process. This leads to the following timeline:

- **Q2–Q3 2026:** Form a coalition of founding library partners willing to commit collections and governance participation. Initiate formal outreach to the Pharos-led Language and Culture Data Lab and relevant AI factory operators. Agree on a coordinating entity and basic governance principles. Begin the pilot, including GRIN ingestion testing and initial VLM re-OCR benchmarking.

- **Q3–Q4 2026:** Define technical specifications and roles for the production system, building on pilot findings. Establish ingestion standards and pathways, agree on output formats, and confirm compute and storage providers. Begin recruitment of core staff.

- **Q4 2026:** Complete the pilot and publish pilot dataset plus pilot findings, including throughput data, cost benchmarks, and quality comparison between Google OCR and VLM re-OCR output. Use these findings to finalise the production system specification and business plan.

- **Q1–Q2 2027:** Launch the EBDC as an operational entity. Begin ingesting and processing collections from founding library partners on a continuous basis.
- **Mid 2027:** Publish the first versioned production dataset releases.

This timeline is aggressive but achievable if the coalition-building phase moves quickly and the technical specifications can be agreed without extended deliberation. The most significant risk to the timeline is not technical but organisational: the time required to align multiple institutions on governance, legal arrangements, and contribution terms. A secondary risk is the speed of GRIN extraction — practical experience suggests that extracting large collections through the GRIN interface can take considerably longer than anticipated, and ingestion testing should begin as early as possible to surface any bottlenecks before they affect the production timeline. Generating a reliable picture of GRIN's practical throughput constraints is one of the primary reasons for conducting the pilot ahead of the production build.

Finally, early clarity on the coordinating entity and governance model (section 4) is a precondition for the operations launch in Q1–Q2 2027.

5.3 Cost structure

The financial pathway from concept to operational infrastructure falls into three stages, each with a distinct cost profile.

The **pilot** is a one-off investment of approximately €75,000 in out-of-pocket costs — roughly €14,500 covering infrastructure and the development of the discovery interface, and €60,000 for coordination, the latter including preparation of the business plan for the production phase. The pilot's purpose is to resolve the single remaining cost variable — the re-OCR model and GPU selection — and to produce that business plan.

The transition to **production** involves two different kinds of cost. The first is a one-time re-OCR compute cost of between €241,000 and €2,361,000 at commercial rates; because it is non-recurring, it is precisely the cost that time-limited AI factory compute is best suited to absorb, potentially reducing it to near zero. The second is the recurring annual cost of running the service — storage rising to approximately €90,000 per year at full scale, together with four to five FTE and ongoing processing — which totals well below €1 million per year. Unlike the one-time compute cost, this recurring base cannot rely on time-limited compute and requires a durable funding model.

Identifying that model, likely combining membership contributions with sponsored resources, is a task for the business planning process that follows this study.

5.3.1 Pilot

Infrastructure costs for the pilot comprise one fixed component — S3 storage with backup and the ingestion staging server, estimated at €3,600 over six months — and one variable component: GPU compute for VLM re-OCR, determined by the model and GPU ultimately selected and addressed through three costed scenarios (conservative, moderate, optimistic). The required OCR model evaluation of three VLMs on a sample of 500 books costs €500–

€950; the optional full re-OCR of the 50,000-book collection, contingent on funding, adds a further €4,000–€39,400. Total pilot infrastructure costs are therefore estimated at €4,100–€4,550 for the required scope, or €8,100–€44,000 including the optional full re-OCR.

Labour costs for the pilot are estimated at approximately €70,000, comprising development of the EBDC-controlled discovery interface (approximately €10,000) and project coordination (approximately €60,000, covering technical coordination by Datovis and overall project coordination by Open Future). These figures are indicative and do not include in-kind contributions from partner institutions.

5.3.2 Production system

Production system infrastructure costs separate into two components: a recurring storage cost and a one-time re-OCR compute cost.

- Holding the full production corpus — scans, OCR text, metadata, and versioned dataset snapshots — on EBDC-controlled S3 infrastructure with offsite backup is estimated at approximately €262,000 cumulatively over a five-year horizon. Storage spend ramps as the corpus is ingested, from a low base in the first year to approximately €90,000 per year once the full corpus is held. This figure is based on the documented GRIN maximum scan size and is therefore a conservative upper bound; actual costs will be lower.

- Re-processing the full corpus with a VLM-based OCR model is a one-time cost, estimated at between €241,000 and €2,361,000 at commercial GPU rates depending on the model and GPU selected — the choice the pilot is designed to resolve. This cost can be reduced to near zero if the compute is sourced from EU AI factories under a data-for-compute arrangement rather than purchased commercially; because it is a one-time rather than recurring cost, time-limited access to AI factory compute is well suited to covering it. Establishing such an arrangement, on terms consistent with the EBDC's governance principles, is a key objective of the pilot.

Personnel costs for the production phase cannot be reliably estimated at this stage. They depend on governance arrangements, the division of responsibilities between the coordinating entity and contributing libraries, and the extent to which operational tasks can be absorbed by existing institutional capacity. Based on the functional requirements of the production system, a realistic estimate is in the range of 4 to 5 FTE covering technical pipeline operations, library relations and ingestion coordination, governance and external relations, and overall programme coordination. The actual cost of these positions will vary considerably depending on where they sit institutionally — whether in a national library, a European foundation, or a dedicated operational entity. Taken together with the infrastructure cost estimates above, however, this suggests that a fully operational EBDC should be achievable at a total annual cost well below €1 million in all scenarios — a figure that compares favourably with the scale and strategic value of the infrastructure being built. The remaining budget variable is the re-OCR model/GPU selection; storage costs are modest in all scenarios. The pilot will narrow these estimates further, enabling a precise budget for the production phase.

Establishing a more precise personnel cost model is identified as an open question to be addressed as part of the business planning process that follows this feasibility study in parallel with the pilot.

6 The opportunity and obligation to act now

The case for building the European Books Data Commons developed above does not rest on opportunity alone. The material conditions for acting are in place: the collections exist, the technology to process them has become accessible, the embargo terms that have restricted access are in the process of being reduced, and the European policy environment is actively seeking exactly the kind of infrastructure the EBDC would provide. What is required now is a decision to act — and for both the libraries that hold these collections and the policymakers who have committed to making European cultural heritage available for AI development, that decision is not optional. **Inaction comes down to a choice to leave the current situation in place: one in which millions of digitised public domain books, produced with public funding and held in public institutions, remain effectively inaccessible while a single commercial operator retains privileged access to the corpus it helped create.**

6.1 For the libraries

European (national) libraries have invested decades and significant public resources in digitising their collections. The Google Books partnerships that drove the bulk of this digitisation were premised on a straightforward exchange: Google provided the scanning infrastructure and returned digital copies to the libraries. Those copies have been held by libraries ever since, largely unused for large-scale re-use, constrained by embargo terms that are now in the process of being reduced to a level the Open Data Directive has always specified.

With that barrier removed, the question libraries face is whether to make these collections available through infrastructure they control, or to allow the current situation to persist. In the current situation libraries are experiencing unmanageable automated scraping of their online collections by actors seeking training data through whatever channels are available. **This demand exists and is not going away; the only question is whether it is met through governed, library-controlled infrastructure or through informal and ungoverned channels that give libraries no visibility, no governance role, and no ability to ensure that the data is used in ways consistent with their public interest mission.**

There is also a narrower institutional argument. Libraries that contributed collections to Google Books received digital copies in return. Those copies are an asset — one whose value has increased substantially as the demand for large-scale book data has grown. Contributing those collections to the EBDC is the mechanism through which that asset generates value for the European public rather than remaining locked in institutional storage.

Beyond the immediate value of contributing existing collections, the EBDC also positions libraries to develop new kinds of bulk data services that they currently cannot offer. The demand for large-scale, well-documented, multilingual book data is not a passing trend — it reflects a structural shift in how knowledge is processed and used, and it is a shift in which

libraries are exceptionally well placed to play a central role. An operational EBDC gives libraries a shared infrastructure on which to develop that role: the ability to offer versioned, provenance-documented, quality-annotated datasets as a service to AI developers and researchers, on terms that libraries set and govern. This is a fundamentally different relationship with the AI ecosystem than the current one, in which libraries are at best passive sources of data scraped without permission and at worst entirely absent from a market that is consuming their collections anyway. Building the EBDC is the mechanism through which European libraries can move from that position and become part of the Public AI infrastructure that underpins AI development in Europe.

The EBDC does not require libraries to take on operational risk or build infrastructure they cannot afford. It asks for collections, governance participation, and a commitment to the principle that public domain material should be publicly available. The infrastructure, the processing, and the distribution are shared costs.

6.2 For policymakers

The European Commission's Data Union Strategy commits to making European cultural heritage collections available for AI development. The AI infrastructure programme is building data labs whose explicit function is to connect data spaces with AI factories. In this context the Commission has set a target of 30 million cultural heritage objects available for the AI ecosystem. What is currently missing is the concrete infrastructure to deliver on this objective.

The EBDC is that infrastructure for European book data. It operationalises the data lab concept for a specific, significant, and currently underserved category of cultural heritage material. It connects the common European data space for cultural heritage to the AI Factories and the AI ecosystem that they serve and that needs multilingual training data. And it does so for a corpus — millions of digitised public domain books in dozens of European languages — that the commercial market cannot replicate and that European AI developers have explicitly identified as a gap.

The policy case for acting now rather than later is straightforward. The Google embargo reduction that is expected in the near term will bring millions of volumes within reach of the EBDC simultaneously. That is a one-time availability event and the value of acting now is precisely that the EBDC can be positioned to receive that material as it becomes available, rather than scrambling to build infrastructure after the fact.

The EBDC is not a large or expensive initiative relative to the infrastructure investments the Commission is already making in AI. Its annual operating costs, once at scale, are well below €1 million. Its potential contribution — a multilingual corpus of millions of European public domain books, governed by the libraries that hold them, available without restrictions to European AI developers and researchers — is disproportionately large relative to that cost. The question is not whether Europe can afford to build it. It is whether Europe can afford not to.

Acknowledgements

This report was written by Paul Keller (Open Future) for the Koninklijke Bibliotheek (KB — National Library of the Netherlands) and the Europeana Foundation. The report has been written under the guidance of a steering committee composed of representatives from the KB and Europeana. Technical input has been provided by Sebastian Majstorovic (Datovis).

The author is grateful to the experts and stakeholders who gave their time and insights for the interviews conducted as part of the study.