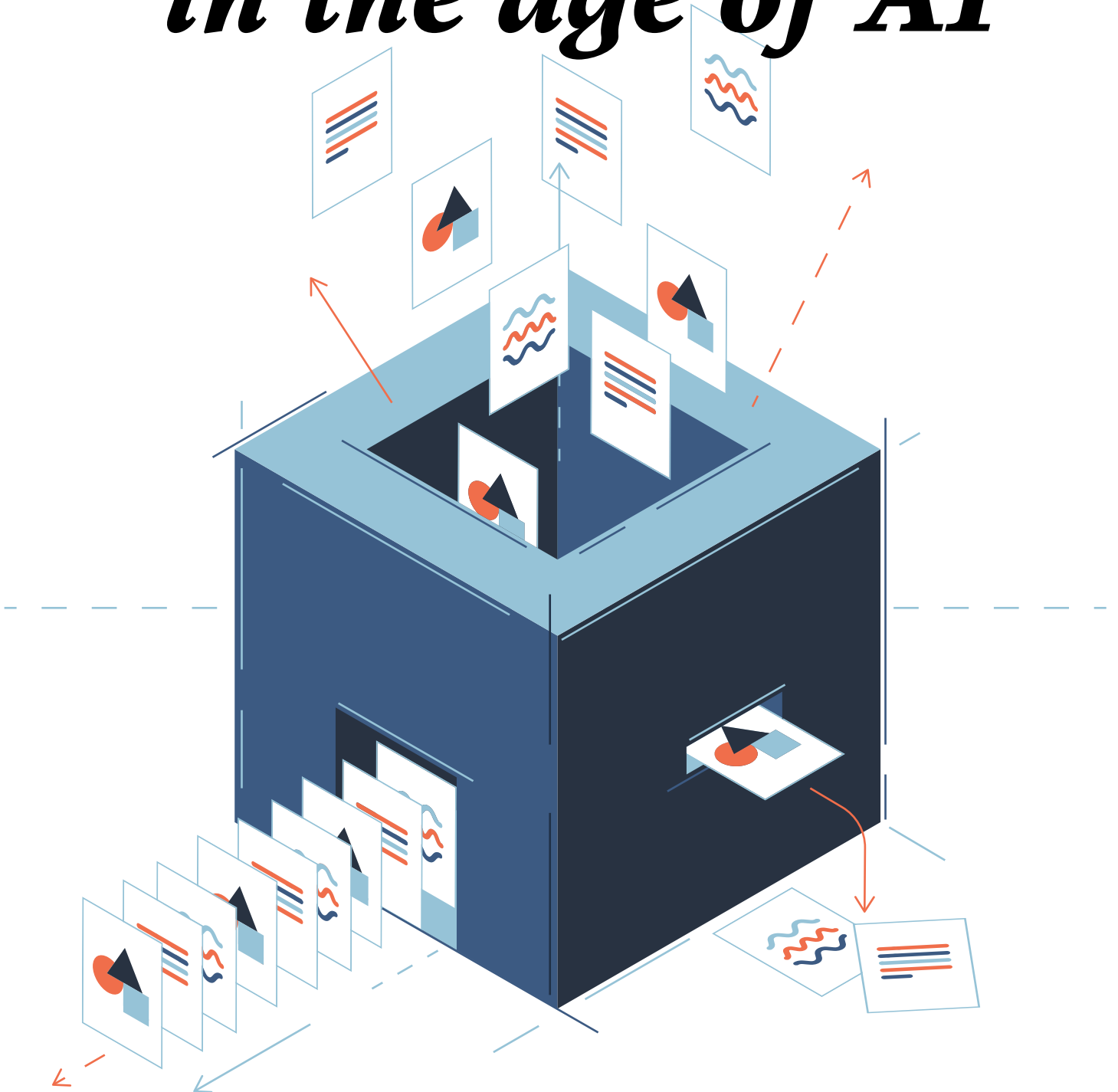


*Impulse paper*

# ***Publishing cultural heritage in the age of AI***



December 2025

# *Preface*

This paper has been commissioned by the Europeana Foundation to the Open Future Foundation, as part of and a contribution to the ongoing [Alignment Assembly on Culture for AI](#). The Alignment Assembly—led by the Europeana Foundation and the Netherlands Institute for Sound and Vision—is a collective intelligence and consultation process that has been taking place within the common European data space for cultural heritage since May 2025. One of the key topics that emerged from this Alignment Assembly concerns the opportunities and challenges of positioning heritage data as responsible AI training material. This paper explores this topic further, focusing on generative AI and its implications for data sharing in the cultural heritage sector. In doing so, it aims to provide further stimulus and input for discussion across the Europeana Initiative and the data space, and further scope one of the main topics identified in the Alignment Assembly exercise.

The paper is organized in two parts: the first outlines the relevant technological and legal context, and the second sets out a proposal for a differentiated access model for cultural heritage data. The model outlined here is not intended as a finished framework, but rather as an invitation for dialogue, reflection, and collective exploration.

## *Introduction*

It has only been three years since the launch of ChatGPT, an event that marks the beginning of another revolution in how we deal with information, knowledge, and cultural artifacts. Since then, debates about the impact of generative AI have gained prominence across the fabric of our economy and society. In the cultural heritage sector, these debates have largely focused on two questions: first, how cultural heritage institutions can integrate AI tools and services into their operations to improve how they deliver on their mission and reach audiences; and second, how their collections are used for training and operating AI models and the services built on them. This paper addresses the latter question and examines how cultural heritage data can be published under conditions increasingly shaped by the demands of AI developers and users.

The [Alignment Assembly on Culture for AI](#)—a participatory process aimed at gathering insights on AI within the common European data space for cultural heritage and the broader heritage sector—revealed interesting observations on this matter. Naturally, the topic of data access emerged as an important theme within this consultation. Heritage institutions manage a lot of data, and AI companies need data to train and improve their models. The opinions captured in the Alignment Assembly reveal a core challenge: there is broad support in the heritage sector that publicly funded heritage data should be made as openly available as possible to support reuse, research, and innovation. On the other hand, privacy concerns<sup>1</sup>, rights management, and the potential use of opt-outs by institutions or rights holders risk limiting access, sometimes even to public-domain materials.

---

<sup>1</sup> While institutions must continue to comply with privacy, data-protection, and personality-rights obligations when publishing any material, this paper focuses on copyright, related rights, and database rights as the primary levers that shape access for AI training (and use by deployed systems).

This paper builds on and dives deeper into these observations. It is motivated by the belief that the common European data space for cultural heritage—the European Union flagship initiative to accelerate the digital transformation of the cultural heritage sector, with the Europeana Initiative at its heart—offers a unique opportunity to address these complex challenges and dilemmas.

Accordingly, the paper develops a framework that supports cultural heritage institutions in deciding whether—and under what conditions—to make collection data available for AI training. The framework aims to help institutions balance their commitment to open access and public information provision with the need to manage new forms of large-scale reuse that come with the rise of AI. In doing so, it seeks to chart a path for how cultural heritage institutions can contribute to a sustainable information ecosystem.

## ***Part 1—Context: AI Use, Legal Framework, and Constraints***

At a high level, artificial intelligence refers to information-processing systems that, for a given set of human-defined objectives, learn from data to generate outputs such as content, predictions, recommendations, or decisions. In this context, “learning from data” refers to encoding correlations, patterns, and other relationships into a model that can then process new inputs to generate outputs in various forms.

Much of the current debate focuses on generative AI models—systems designed to produce synthetic content—and in particular on Large Language Models (LLMs). However, the underlying principle of training models on data can be applied to many different types of data and used to produce a wide range of outputs.

As a result, all information held by cultural heritage institutions can, in principle, serve as training data for AI models. This includes both the descriptive metadata that makes collections discoverable and the digitized or born-digital cultural heritage materials themselves. Both types of data encode a vast richness of human and societal knowledge and cultural expression. As such, they are extremely valuable for training AI models that aim to mimic and augment human understanding and expression. This explains why AI developers are seeking access to the data held by cultural heritage institutions. This data is valued not only for volume but also for its curated, authoritative quality—distinguishing it from general web data.

Once a model has been trained and fine-tuned, it can be deployed as part of an AI system. Such systems take many forms, including chatbots, image and video generators, and analytic tools. Importantly, not all uses involve the creation of synthetic content. Models are also deployed to analyze, classify, or otherwise process large volumes of information.

This means that trained AI systems can interact with cultural heritage in several ways. First, cultural heritage data can be used for grounding (as additional information that improves the relevance or accuracy of model outputs in response to user input). Second, it can also be provided as reference input (for example, enabling style transfer, where the model learns to reproduce the visual or textual style of specific works, or for generating new, derivative

outputs inspired by or resembling those works). Finally, trained AI systems can operate directly on cultural heritage data to summarize, compare, or otherwise analyze cultural artifacts and the information they contain.

In both cases (training and use by deployed systems), broad access to cultural heritage collections can provide substantial value to the companies developing and deploying AI models. However, the dynamics differ significantly between these two phases.

1. **Training.** Model developers generally require access to large volumes of high-quality data. Once acquired, this data is typically further processed internally—cleaned, filtered, and formatted—before being used for training. Data obtained for training is usually collected once but can then be reused for multiple training runs and model iterations. The primary mechanism for obtaining cultural heritage data is large-scale web crawling, in which automated agents access publicly available online resources and copy them into internal systems, where the material is aggregated into training datasets<sup>2</sup>. This crawling is conducted both by AI developers themselves and by specialized data companies that compile training datasets and either sell them or make them publicly available. Because training data does not need to be delivered in real time, developers also obtain data by other means, such as digitizing analogue books and cultural heritage assets, with the explicit goal of building training corpora. In some cases, developers seek access to cultural heritage data that is not directly available online—either because access is restricted or because it resides in systems not exposed to the public internet. In both cases, access typically requires contractual agreements or partnerships between cultural heritage institutions and either AI developers or specialized dataset providers.
2. **Use by deployed systems.** By contrast, most uses of cultural heritage data by trained and deployed AI systems require real-time access via the public internet. Modern AI systems can retrieve online resources to augment their inputs. This may be user-directed (for example, a request to summarize or analyze specific documents) or autonomous (where the system accesses resources to ground a response or to obtain information that is then included in a response to a query).

From the perspective of cultural heritage institutions, it is often difficult to distinguish between automated requests from crawlers collecting training data and automated requests from deployed AI systems retrieving resources to answer queries. Both show up as “bots” in the logfiles of their websites.

While much of the current debate focuses on the use of cultural heritage data as training data, the impact of requests generated by deployed systems can be expected to grow as such tools become more widely used. Unlike training, which requires data to be collected only once, deployed systems may access the same resources repeatedly, leading to ongoing and potentially significant demand.

---

<sup>2</sup> Other training sources include proprietary platform/user data and corpora acquired via licensing. These are outside the scope of this section, which focuses on cultural heritage data.

**This also means that restricting access by automated systems or bots to online cultural heritage data will not only affect the use of the data for the purpose of training AI models but will also impact the ability of anyone using AI systems to find and analyze the same data.** While there are ongoing efforts to develop standards for communicating more fine-grained preference signals that would allow institutions to distinguish between training and inference-time uses, such standards are not yet widely implemented in production AI systems.

## *The Legal and Technical Framework for Managing Automated Access*

While there are endless discussions about the relationship between the training and use of generative AI systems and copyright law, copyright plays only a secondary role in making cultural heritage collections available for AI training and use. Two overlapping characteristics explain why.

### **Copyright**

From a copyright perspective, there are broadly two types of cultural heritage resources: those still in copyright and those in the public domain. Making available in-copyright works requires permission from the rightholders, while public domain works can be made available online at the discretion of the cultural heritage institutions that hold them. **Once made publicly available, public domain works can be freely used by anyone for any purpose, including training AI models or serving as inputs for deployed AI systems.**

Under European law, lawfully accessible works that are in copyright may be used for text and data mining (TDM) for training AI systems and for automated analysis by deployed systems, subject to the conditions of the TDM exceptions. Both types of uses are instances of TDM. The 2019 Copyright in the Digital Single Market (CDSM) Directive provides two relevant exceptions: Article 3 establishes a mandatory exception for TDM carried out for scientific research by research organizations and cultural heritage institutions with lawful access, while Article 4 establishes a broader exception that allows TDM by anyone for any purpose on lawfully accessible works, unless their rightholders have explicitly reserved this right ('opted-out') in a machine-readable manner.

This means that copyright only comes into play when cultural heritage institutions make in-copyright works available. In order to do this, they must have obtained permission from the rightholders or from organizations such as collective management organizations that issue licenses on their behalf. This also means that Article 4 opt-outs, which need to be made by rightholders, cannot be made by cultural heritage institutions at their own discretion but must be based on explicit requests by rightholders or organizations acting on their behalf<sup>3</sup>.

This leaves cultural heritage institutions with very limited means to invoke copyright to prevent the use of works that they make available online to train AI models or have them used by deployed AI systems:

---

<sup>3</sup> There is some discussion within the cultural heritage community if CHIs issuing licenses on behalf of rightholders can legitimately reserve the rights in question. The author is of the opinion that they can.

1. They must express machine-readable opt-outs for publicly available online content when requested or required to do so by rightholders who have authorized the making available of in-copyright works.
2. They can make machine-readable rights reservations at their own discretion when publishing works for which they are the rightholders themselves (which is relatively rare in the cultural heritage context).

## Database rights

In addition to copyright, cultural heritage institutions may also rely on the *sui generis* database right (SGDR), which protects collections as structured datasets rather than the individual works they contain. Article 4 expressly covers database rights, so database rightholders can also exercise the Article 4 opt-out. This allows cultural heritage institutions to register opt-outs based on database rights that they hold in their digital collections.

Since the database right operates independently of the copyright status of works contained in a database—but applies to the act of extraction of items from the database—such opt-outs can, in theory, be registered for collections consisting of public-domain works as well as those consisting of in-copyright works. While the SGDR offers, in principle, one of the few legal mechanisms that cultural heritage institutions can exercise directly, its use in this context raises significant normative concerns. Applying the SGDR to restrict large-scale reuse of public-domain collections risks reintroducing forms of enclosure that run counter to long-established public-domain and open-access norms. For institutions whose mission is to preserve and expand the cultural commons, relying on database rights to constrain reuse is therefore neither desirable nor structurally aligned with the values that underpin open access to cultural heritage data. For these reasons, the differentiated access model proposed in Part Two of this paper does not depend on asserting database rights, but instead on shaping access conditions and delivery mechanisms in ways that support openness while ensuring sustainability under conditions of large-scale AI reuse.

## Practical Constraints on Enforcing Preferences

While the EU regulatory framework is clear—works for which the rightholders have made a machine-readable rights reservation cannot be used to train AI models—the real-world impact has so far been limited. This is the result of two issues: there are currently no widely used and respected standards for machine-readable opt-outs, and the most widely used mechanism (`robots.txt`) is not designed to handle AI-related opt-outs. More importantly, there is significant uncertainty around the territorial application of the EU copyright rules for entities operating outside of the EU.

**In practice, this means that cultural heritage institutions should assume that opting out of TDM is currently not an effective way to prevent the use of materials they make available for training AI models or for use by deployed AI systems—even in the limited cases where they hold the rights needed to invoke such an opt-out.**

**Conversely, it seems prudent to assume that everything they make available online (or have made available in the past) will find its way into AI training datasets and will likely be used by deployed AI systems.**

Outside of the realm of copyright, there are a number of technological solutions that cultural heritage institutions can use to control automated access to their online resources. These measures are not legal restrictions but technical barriers that aim to limit or manage large-scale scraping by AI developers.

One example is the service offered by Cloudflare, a widely used content delivery and security provider. Cloudflare allows its customers to automatically block a range of known AI scraping bots from accessing any resources on their websites. In addition, Cloudflare is also experimenting with a pay-per-crawl model that would allow crawling in exchange for financial payments to the site operator.

There are also open-source tools that offer similar functionality. One example is [Anubis](#), an open-source software project designed to identify and block automated agents associated with AI dataset harvesting. Unlike commercial services, open-source tools can be deployed directly by institutions on their own infrastructure, giving them more control and independence.

In theory, such technological measures can help cultural heritage institutions manage automated access, but their effectiveness remains unproven, especially against determined scraping efforts that disguise themselves as regular user traffic.

## *Problem statement*

Confronted with the emergence of AI as a new socio-cultural technology, cultural heritage institutions face a strategic choice: whether to encourage, deter, or condition the reuse of their online collections as training inputs for AI systems. This issue surfaced prominently through the Alignment Assembly on Culture and AI, where participants flagged “turning heritage data into training datasets” as a central concern, asking who should be able to use these materials and on what terms.

The re-emergence of how to deal with access to digitized collections is the result of a major shift in the operational context of cultural heritage institutions: When large-scale digitization began, access models and agreements were designed for human-scale use: discovery and consultation of individual works. Today, AI has created demand for collection-level access at an industrial scale. Institutions report rising requests from researchers, innovators, and AI developers for bulk access to digitized collections, often supported by policy narratives that demand access to “high-quality data” for AI. The same infrastructure that enables open public access now enables automated harvesting. This change alters costs, incentives, and the visibility of institutions as authoritative sources. This shift also raises normative questions that are not merely technical: should automated machine access be treated as equivalent to human access—and if not, on what terms?

As discussed above, both legal and technical levers exist but are limited in practice. **Consequently, the key issue for institutions is not only what is legally possible, but what**

**aligns with their public mission, stewardship responsibilities, and sustainability.** In the following section we will develop a public-interest framework that helps institutions decide if and how they should make collections available for AI training and use.

To navigate these challenges, cultural heritage institutions require a structured way to evaluate whether, and under what conditions, to make their collections available for AI training and for use by deployed systems. The next part of the paper proposes such a framework. It begins by identifying the public-interest principles that should guide institutional decisions, and then applies these principles to assess the available access options.

## *Part 2—Proposal for a Public-Interest Framework for Access Decisions*

Most cultural heritage institutions in Europe are publicly funded institutions with a mandate to preserve, steward, and provide access to the works and associated data in their collections. They operate within a broader information ecosystem that supports education, research, historical awareness, and democratic accountability. As AI creates demand for collection-level access—both for training models and for use by deployed systems— institutions need clear, mission-aligned criteria for deciding whether, and under what conditions, to enable such access. This section sets out a public-interest framework: the principles that should guide those decisions.

- **Open access to knowledge:** The overarching principle guiding cultural heritage institutions. It builds on access and reuse policies developed over the past two decades and codified by institutions (e.g., the Europeana Public Domain Charter), governments, and funding bodies. This principle applies within the boundaries of the copyright framework, which means institutions must work with rightholders when making in-copyright collections available, while committing to making public domain materials as openly accessible as possible.
- **Equitable, non-discriminatory access and non-exclusivity:** Access conditions should be transparent, proportionate, and applied on a non-discriminatory basis. Institutions should avoid exclusive arrangements that foreclose public or competitive access and ensure that terms do not, in effect, exclude public-interest users or smaller actors.
- **Trustworthiness, authenticity, and authority:** Cultural heritage institutions uniquely preserve the historical and cultural record and provide reliable context and provenance. In the digital environment—and given AI’s propensity to hallucinate or amplify misinformation—their role as authoritative, trustworthy sources is even more important, including safeguarding the meaning and context of their collections.
- **Support for research, education, and innovation:** Enabling these activities is a core public-interest justification for open access policies and for public funding of cultural heritage institutions. Downstream use in research, educational, and other contexts is one of the mechanisms by which institutions generate economic value for society at large.

- **Economic sustainability:** Institutions must remain economically sustainable and allocate resources in line with their public mission. In the context of AI, rising demand for bulk access to digitized collections creates financial and operational burdens; access decisions should account for these costs so openness remains viable.

These five principles map onto the [three core values that underpin Europeana’s mission](#): Usable, Mutual, and Reliable. Usable encompasses the commitment to open and equitable access and the support for downstream research and innovation. Mutual highlights the need for fair, reciprocal arrangements that ensure shared benefit and support economic sustainability. Reliable addresses the issues of trustworthiness, authenticity, and authority.

As illustrated by the responses to the Alignment Assembly and subsequent discussions within the Europeana Initiative and the common European data space for cultural heritage, generative AI brings these principles and values back into sharp focus. **As institutions set policies for making collections and associated data available for training AI models and for use by deployed AI systems, they must reconcile a new, intense demand for collection-level access with principles shaped in an era of supply-side efforts to stimulate reuse.**

The next section translates this tension into concrete options and assesses their consequences against the public-interest framework.

## *Assessing Access Conditions and Modes*

This section sets out a practical framework for making cultural heritage data available online, starting from a realistic baseline: any material made publicly accessible is likely to be used for training and accessed by deployed AI systems. The framework considers collections and their associated metadata together, reflecting demand for both. It focuses on materials that institutions can lawfully publish—public-domain items or works for which permission has been obtained—and notes additional considerations for in-copyright materials at the end of the section. Options are assessed against the public-interest principles defined above.

### **Axis 1—Access conditions (open → closed)**

The x-axis on the matrix below distinguishes four different sets of access conditions, ranging from open to fully closed. These are:

- **Open:** Public access with no additional provider-imposed restrictions beyond the legal framework. Access is, in principle, anonymous and non-discriminatory. Only basic anti-abuse measures (e.g., rate limiting).
- **Controlled:** Access available to anyone under published, uniform, non-exclusive terms (licensing, API-key, or click-through), with transparent rate/volume limits and access logging.
- **Conditional:** Access is granted to users or for uses that meet published criteria (e.g., accredited research organizations, public-interest projects). Other types of uses—including large-scale or commercial uses—may require case-by-case approval and

may be made conditional on specific terms, including payment of usage fees and reporting obligations.

- **Closed:** No online access (internal use only). This includes on-demand access and internal forms of access, such as reading-room access.

## Axis 2—Access modes (how data is delivered)

The y-axis on the matrix below distinguishes three different access modalities:

- **Individual item access:** Human-scale webpages, viewers, per-item downloads (incl., images, PDFs).
- **Programmatic (API) access:** Query/endpoint access to items and/or metadata—suitable for real-time uses by deployed systems (e.g., REST/GraphQL/SPARQL, IIIF endpoints, or MCP for AI assistants).
- **Bulk / collection-level access:** Data dumps/snapshots, batched exports, or object-store access optimized for high-throughput reuse (useful for AI training and other research workflows).

## Access conditions vs modes

The access matrix below illustrates how different combinations of access conditions (x-axis) and access modes (y-axis) align with the core principles identified above. It highlights the central tension between maintaining open and usable access (OA) and ensuring economic sustainability (ES), and shows how this tension affects equitable, non-discriminatory access (EQ). As the access modes shift from individual item access to API access and finally to bulk or collection-level reuse, the matrix makes clear that access conditions need to become correspondingly more structured. The shading indicates where a particular combination supports these principles and where it creates friction, with stronger colours signalling a clearer match or conflict:

Access condition → ↓ Access mode	Open	Controlled	Conditional	Closed
Individual	OA✓ EQ✓ ES✓	OA~ EQ✓ ES✓	OA× EQ~ ES✓	OA× EQ× ES~
API	OA✓ EQ✓ ES~	OA✓ EQ✓ ES✓	OA~ EQ~ ES✓	OA× EQ× ES~
Bulk	OA✓ EQ✓ ES×	OA~ EQ✓ ES~	OA~ EQ~ ES✓	OA× EQ× ES✓

This matrix reveals key insights regarding how cultural heritage institutions should approach publishing collections data in the age of AI in ways aligned with their public-interest missions.

For individual item access and for programmatic (API) access, this analysis points to a clear best fit: fully open access for individual items and controlled access for APIs. Bulk access is different. While the matrix indicates that bulk access could, in principle, be offered under either controlled or conditional terms, conditional access provides a materially better fit with institutional sustainability and stewardship responsibilities grounded in the value of mutuality.

Bulk reuse generates the highest operational costs (assuming that measures like rate-limiting can properly shield individual and API access from abuse) and is the point at which asymmetries between public institutions and large-scale (commercial) users become most pronounced. Controlled access, based on uniform terms, rate limits and logging, offers only limited means to manage these pressures and, in practice, tends to amount to quasi-open bulk access. By contrast, conditional access allows institutions to differentiate between non-commercial and large-scale commercial uses and to ensure that bulk users participate in the maintenance of the commons. Because ordinary public access remains open at the individual-item level and non-discriminatory for API use, the impact on equitable access is limited. For these reasons, conditional access is the approach that most effectively reconciles open access commitments with economic sustainability in the context of bulk or collection-level reuse.

More structurally, the matrix shows that—despite the profound changes brought by generative AI—openness, to a degree, should remain a key principle when making collections data available. Closing access to existing (or new) digital collections is not a structurally sound response to the challenges posed by generative AI. Fully closing access has overwhelmingly negative consequences that undermine core principles underpinning the work of cultural heritage institutions in the digital sphere. At the same time, the matrix also indicates that, when it comes to making data available in bulk for AI training, doing so under fully open conditions does not align well with these principles.

Finally, the matrix shows that as the access modes place greater demands on the infrastructure and the commons—moving from individual item access to programmatic (API) access and, finally, to bulk or collection-level access—the access conditions should become correspondingly more structured: open for individual item access; controlled for programmatic (API) access; and conditional (eligibility-based) where clearly justified by published criteria.

### *A differentiated access model*

In this section, we take a closer look at the preferred option for each of the three access modes.

For **individual item access**, the preferred option remains **fully open access** to digitized collections (both metadata and, where possible, the digital objects themselves). This aligns with established practice, and the emergence of AI does not, on its own, justify imposing new restrictions or conditionalities at this level.

At the same time, AI has increased automated demand—notably through web scraping and crawling—which can be costly to manage. To address this, cultural heritage institutions

should steer high-volume or automated uses toward the other two access modes: programmatic (API) access for use by deployed AI systems and bulk/collection-level access for the acquisition of training data. Institutions should clearly signal and document the availability of these access modes. In addition, they may apply proportionate traffic-management measures (e.g., rate limiting, throttling, or blocking unidentified automated traffic) to protect individual item access without undermining openness for human users.

For **programmatic (API) access**, the preferred option is **controlled access**. This mirrors existing practice: many institutions provide programmatic access to collection data through APIs subject to lightweight terms of use, issuance of API keys, transparent rate limits, and basic logging. Such arrangements let institutions understand who is accessing their collections via programmatic interfaces, manage capacity through quotas, and enforce rules against abuse or excessive use. This can also help steer high-volume, AI-training-related uses towards mechanisms designed for bulk or collection-level access. It is also desirable to steer AI systems that need real-time access toward these programmatic interfaces rather than scraping individual item pages. Over time, new standards for AI clients (particularly so-called "agentic" systems that retrieve information on behalf of users) are likely to emerge, which institutions can adopt as they mature.

For **bulk / collection-level access**, the preferred option is **conditional access**. This type of access is not yet widely used within cultural heritage contexts. Its key difference from the previous two access regimes is that it allows segmentation between different categories of uses or users. In practice, only a small subset of users require bulk access, including—but not limited to—those seeking collection data for AI training<sup>4</sup>.

Existing examples of conditional access include the release of the [institutional books corpus](#) by the Institutional Data Initiative, which makes a collection of 983,000 public-domain books available in bulk via Hugging Face [under terms that allow non-commercial use only](#). The IDI has also encouraged entities interested in commercial (including AI-training) uses to negotiate bespoke terms that include monetary contributions. This "gated access" model shows that, at least for high-quality collections, it can be feasible to require bulk users to contribute back to the institutions making the data available.

This model addresses several AI-related challenges around collection-level access. First, by releasing large-scale datasets in bulk using formats and interfaces optimized for AI, cultural heritage institutions position themselves as active participants in AI development. By publishing data directly, under their control and on their terms, they maintain visibility as trusted, authoritative sources. Clear differentiation between free use for research and other non-commercial purposes and paid use for commercial purposes supports alignment with public-interest objectives. If adopted at sufficient scale, paid commercial use can create a mechanism for commercial AI developers to contribute back to the information commons,

---

<sup>4</sup> An additional issue in current discussions concerns "no-resharing" conditions for bulk datasets, used by initiatives such as the Institutional Data Initiative or Mozilla's Data Collective. Such conditions are designed to prevent immediate re-hosting of bulk collections and to ensure that high-volume users access data through the dedicated, conditional interface. In the context of a differentiated access model, prohibiting unrestricted bulk re-sharing should not be seen as a restriction on openness at the level of cultural heritage access; it is a practical requirement for maintaining the integrity and sustainability of conditional bulk access offerings.

rather than offloading scraping costs onto publicly funded institutions. Finally, steering bulk users to dedicated access channels can reduce load on individual-item access and lower operational costs.

## *Normative concerns*

Beyond these operational considerations, limiting openness at the level of bulk access also reflects a normative concern. Current dynamics of AI training and deployment [concentrate control over access to and value derived from cultural data in a small number of highly resourced actors](#). Unrestricted bulk reuse by such actors risks turning the cultural commons into a one-way input for private model development, undermining the visibility, sustainability, and public mission of the institutions that maintain it. Introducing conditional access for large-scale, commercial uses is therefore not a restriction of the public domain as such, but a measure to preserve its public value under conditions of structural asymmetry.

This approach rests on several assumptions that will need to be tested in practice. It requires that institutions offering bulk-level access can effectively steer bulk users toward dedicated interfaces. This may require technical means to limit large-scale access via individual-item pages and/or strong incentives for bulk users to do “the right thing”. The overall economic feasibility will also likely require aggregation of data into large-scale, high-quality datasets.

In terms of economic sustainability, such efforts should not be positioned as attempts to create additional revenue, but rather as contributions back to the commons that allow institutions to carry out their missions. The objective is that economic returns help cover the costs of digitization and data preparation, subsidising free use for research and other non-commercial purposes.

It is important to recognize that this model breaks with some established norms around copyright, public-sector information and open data. By requiring some types of users to pay for access to bulk collections optimised for AI use, it departs from the longstanding practice that access to public-domain works or data should be provided for free and under non-discriminatory terms. At the same time, conditional bulk access does not assert new intellectual property rights in public-domain reproductions; it governs delivery at scale while individual and API-level access remain open and non-discriminatory.

In this sense the proposal mirrors developments taking place elsewhere in the open ecosystem. The Wikimedia Foundation—confronted with increasing costs caused by AI-related scraping of its projects—has formulated the principle that “[Our content is free, our infrastructure is not](#)” Wikimedia has for some time operated a differentiated access model through the Wikimedia Enterprise project, which provides access optimized for the needs of large-scale users of Wikimedia content in exchange for payment. This illustrates that introducing structured access pathways for high-volume commercial users is not a departure from openness, but a pragmatic response to sustainability challenges faced across the commons<sup>5</sup>.

---

<sup>5</sup> In the 2024-25 fiscal year, [Wikimedia Enterprise generated a net positive contribution to the Wikimedia Foundation’s budget for the first time](#). This illustrates that steering well-resourced, large-scale users toward specialised access services can produce meaningful financial returns that support the sustainability of open infrastructures.

At the same time, the wider EU policy context is also shifting. As part of the proposed merger of the Open Data Directive into the Data Act, the European Commission has introduced a derogation from the Directive's general non-discrimination principle for very large enterprises (in practice, the gatekeepers designated under the Digital Markets Act). This signals an emerging recognition that, under conditions of structural asymmetry in digital markets, uniformly open and non-differentiated access may no longer be sufficient to protect the sustainability of public-sector information resources.

This proposal reflects a growing acknowledgment that open access must be balanced against the disproportionate burdens placed on public infrastructures by a small number of highly resourced commercial actors. These changes—if adopted—would strengthen the case for a differentiated access model for digitized cultural heritage: a sector-specific, mission-aligned approach that operationalizes the same concern by ensuring that high-volume, commercial-AI uses contribute to the sustainability of the commons.

Finally, it should be recalled that this differentiated access model has been developed on the assumption that the collection of data in question is either in the public domain or cleared for open access. While nothing structurally prevents the extension of this approach to collections that include in-copyright works, this will require additional refinements, particularly concerning the distribution of revenues generated by conditional access to bulk collections.

## *Invitation for Feedback*

A transition toward more differentiated forms of access to cultural heritage data will require careful deliberation within the cultural heritage community. The model proposed in this paper is intended as a starting point for these discussions, rather than a prescriptive blueprint. Its feasibility and desirability will ultimately depend on collective assessment by institutions across the Europeana Initiative, the common European data space for cultural heritage, policy makers, and the broader sector. The Europeana Foundation and Open Future, therefore, invite feedback on the approach presented here, with the aim of informing future work toward a shared framework for conditional access that aligns with the mission, values, and sustainability needs of cultural heritage institutions. Readers are encouraged to reflect on the ideas put forward in this Impulse Paper, as we will be gathering feedback and insights in the coming months.

## *Acknowledgements*

This paper was commissioned by the Europeana Foundation to Open Future and written by Paul Keller. It is part of, and a contribution to, the ongoing [Alignment Assembly on Culture for AI](#)—a collective intelligence and consultation process that has been taking place within the common European data space for cultural heritage since May 2025. We thank Alek Tarkowski, Antoine Isaac, Ariadna Matas, Harry Verwayen, and Lorena Aldana for their contributions and feedback.

[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.



This report is published under the terms of the [Creative Commons Attribution License](#).