Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cooperate And Compete: Archives Can Gain More By Expecting Less From AI Training

Jinfang Niu

jinfang@usf.edu

University of South Florida

Research Article

Keywords: AI model training, archives, archival provenance, data provenance, item-level provenance

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11392603/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cooperate And Compete: Archives Can Gain More By Expecting Less From AI Training

Abstract

AI companies have increasingly approached archival organizations seeking access to unpublished or undigitized collections for model training. Many archival organizations have responded cautiously to these requests and lean toward rejection due to various concerns. Two major concerns are about the loss of item-level provenance in AI outputs and the loss of control over archival knowledge absorbed into AI models. This paper examines these two concerns and challenges some of the assumptions underlying them. It argues that the loss of item-level provenance in AI outputs may in fact constitute a competitive advantage for archival organizations. It analyzes the potential costs of refusing to work with AI companies and suggests that archival organizations negotiate agreements that preserve their competitive advantage while indemnifying them against liability arising from the misuse of archival knowledge incorporated into AI models. The paper further suggests that archival organizations tailor policies according to the legal status of individual collections, rely on RAG technology for reliable item-level citation, and advocate for government policies that encourage large AI companies to contribute to public goods. Such policies could help improve the public image of AI companies and motivate more archival organizations to collaborate with them.

Keywords

AI model training, archives, archival provenance, data provenance, item-level provenance

Introduction

Having trained their models on vast quantities of content from the open web, AI companies have increasingly approached archival organizations seeking access to unpublished or undigitized collections for model training (Trust in Archives Initiative 2026). Many archival organizations have responded cautiously to these requests and lean toward rejection. For example, the Rockefeller Archive Center (RAC) (2026), a well-funded private archival institution, states: "We never sell or make deals with AI companies to use our archival holdings as training data for LLMs." In fact, RAC currently does not even use AI to facilitate its archival work. RAC cites concerns about the legal, ethical, environmental, and human costs associated with AI technologies. More specifically, some of its collections are subject to access restrictions arising from privacy concerns, copyright limitations, or donor agreements, and it does not wish to expose those materials to AI crawling. RAC also points to the environmental impacts of AI, including substantial resource consumption and pollution, as well as the potential for AI technologies to displace human labor.

The Trust in Archives Initiative (TAI) (2026a) is a coalition that includes the Association of Moving Image Archivists (AMIA), the Archival Producers Alliance (APA), the Society of American Archivists (SAA), and other organizations associated with audiovisual archives. Its licensing templates state that "the Licensor [archival organization] does not permit the use of licensed Content, or any materials from its Content library, or associated metadata … for training AI software, large or small language models, or other AI software tools." Among TAI's concerns are the risks of reputational damage and financial loss arising from misuse of archival content. For example, if archival media are used in AI training, resulting models could generate harmful content

that negatively affects an archive's reputation or produce derivative works that closely resemble original media, potentially undermining licensing revenues and diminishing the value of the original materials (TAI 2026b).

The University of Virginia Library (UVA) has published an archival AI protocol for responding to AI-training requests. The protocol states that, by default, "training or fine-tuning of general-purpose commercial foundation models on archival collections are not permitted" (Lo 2026). UVA further coordinated a group of research libraries to develop the Statement of Shared Practice: AI Training Requests and Unique Cultural Collections (University of Virginia Library 2026). As of April 2026, the statement had been signed and endorsed by eleven archival organizations. It declares that, "For broad commercial training where provenance and meaningful control are not realistic, each institution will independently apply heightened scrutiny and a presumption against approval" (University of Washington Libraries 2026). These research libraries identify two primary concerns. The first is the loss of item-level provenance. Contemporary AI models generally do not preserve explicit links between outputs and the specific records from which information was learned. Without such provenance, repositories may lose opportunities for attribution, while users are unable to verify the origins of AI-generated content. The second concern is the loss of control. Once archival materials have been absorbed into model weights, repositories cannot easily remove them or control their downstream uses. Some archival organizations would prefer to retain the ability to halt the ingestion of their materials and, where necessary, require the decommissioning or destruction of models trained on their collections (University of Washington Libraries 2026; Lo 2026).

Despite this strong language opposing commercial AI training, both TAI and UVA have indicated a willingness to engage with AI companies and negotiate appropriate terms. For example, TAI has developed a toolkit to help archival organizations evaluate potential collaborations with AI companies (TAI 2026b). Revenue sharing from AI-generated products is one of the headline considerations identified in that toolkit. Similarly, the UVA archival AI protocol states that "the archival organization retains the right to seek financial compensation, cost recovery, or in-kind benefits for AI-related uses of its archival materials, especially in commercial or revenue-generating contexts" (Lo 2026, p 11). In addition, the Statement of Shared Practice explicitly acknowledges its temporary nature and indicates that it will be reviewed after one year. TAI (2026c) likewise notes its willingness to revise and adapt its licensing language to remain practical, responsive, and aligned with community needs.

Looking ahead, increasing numbers of archival organizations are likely to receive similar requests, either directly from AI developers or indirectly through data suppliers acting on their behalf. These organizations will therefore need to make informed decisions that reflect their missions and best interests. This paper examines the major concerns that archival organizations have regarding AI training on their collections and proposes strategies through which they may benefit more by expecting less from AI training. Ultimately, it seeks to alleviate some of these concerns while helping archival organizations adapt or reformulate their AI-training policies.

About item-level provenance

Archival provenance

The SAA Dictionary (https://dictionary.archivists.org/) defines archival provenance as "information regarding the origins, custody, and ownership of an item or collection." Archivists use the concept of provenance as a foundation for arranging and describing archival materials: records sharing the same provenance are kept together and separated from records created by other individuals, families, or organizations. Provenance information also provides context and helps users understand the content of archival records and evaluate their authenticity. In principle, this provenance information, like other forms of information, can be learned and retained by AI models if it is included in the training data.

The item-level provenance sought by some archival organizations for AI-generated outputs differs from the archival provenance defined by SAA. It refers to the specific archival records used in generating a particular AI output. The UVA Archival AI Protocol states that the minimum citation requirements for AI outputs should include the archival repository, the relevant archival collection, and the specific item within that collection. This mirrors traditional archival citation practices, in which scholars are expected to provide item-level citations for archival records used in their publications. Such citations enable users and archivists to locate the underlying record, assess its authenticity, and attribute credit to both the archival repository and the creator of the original records.

However, when archival records are used for AI training, they are first transformed into machine-readable representations and then incorporated into model weights. These model weights are no longer archival records, and an AI model is not an archival repository. In some respects, an AI model resembles a human memory system. When a person writes an article, we do not expect to know the specific memories or cognitive processes used to produce that article, nor do we expect to identify every source that contributed to those memories. Instead, we cite the article and its author, thereby attributing credit and assigning responsibility for the work. Similarly, when an AI model generates an output, it may be unreasonable to expect identification of the specific model parameters involved in generating that output or of every source document that contributed to the formation of those parameters. A more reasonable expectation is to cite the model and the prompt used to generate the output. Because models evolve through retraining and may produce different outputs depending on the user or agent interacting with them, citations should also record when the output was generated and by whom.

Traditionally, attribution for the use of archival collections has been a common expectation of archival organizations. However, attribution is not always required at the item level. For example, the National Archives and Records Administration (NARA) expects item-level citations when its records are used in manuscripts, scholarly journals, online publications, websites, and books. However, for audiovisual productions using NARA's special media materials, the agency requests only repository-level attribution:

> "To credit the use of NARA special media material obtained from our holdings in an audiovisual production, such as a film, broadcast documentary, or online presentation, we ask that you provide the general credit line: Courtesy: National Archives and Records Administration." (NARA 2023)

In addition, creative and transformative uses of archival materials often do not require formal attribution. For example, a writer who consults archival materials and subsequently produces a novel or a poem is generally not expected to provide citations to every archival source that informed the creative work.

Data provenance

In the AI context, data provenance refers to the comprehensive documentation of a dataset's origins and lifecycle, including its sources and collection methods, transformation processes and methodologies, ownership history, quality-assurance procedures, usage permissions, licenses, restrictions, and modification history (Data Foundation 2025). Multiple stakeholders participate in the lifecycle of AI training data, including content creators, dataset creators, AI developers, and model users. Data provenance helps AI developers understand their datasets and evaluate content authenticity, quality, diversity, bias, copyright status, privacy implications, and intended uses (e.g., non-commercial use), all of which are critical for responsible model development. It also supports content creators and dataset creators in seeking attribution and compensation while helping users understand the capabilities and limitations of AI systems (Longpre et al. 2024a).

When archival materials are used for AI model training, data provenance encompasses the entire lifecycle of the training data, including the creation of archival records, archival custody and control, the transformation of records into machine-readable representations, and their incorporation into AI models. In addition to the stakeholders identified above, archival organizations that steward archival collections become important participants in this lifecycle. Archival provenance focuses primarily on the origins, ownership, and custody of records up to the point of archival control. Data provenance, by contrast, extends beyond archival provenance by documenting how archival materials are acquired, processed, transformed, and ultimately incorporated into datasets used for AI training. In this sense, data provenance is broader in scope than archival provenance.

Increasingly, regulators and lawmakers are requiring AI developers to disclose aspects of data provenance in order to improve transparency, accountability, and trustworthiness in AI systems (Longpre et al. 2024a). For example, Article 53(1)(d) of the European Union's AI Act requires providers of general-purpose AI models to publish a summary of the content used to train their models (European Commission, 2025). At the same time, these requirements explicitly balance transparency with the protection of trade secrets and confidential business information. Recital 107 of the AI Act states that the published information should be “generally comprehensive” but not “technically detailed,” and that it should take the form of a summary rather than a detailed disclosure of the specific data and works used in model training (European Commission 2025, p 2).

The specific information required by the European Commission's compliance template includes dataset modality (e.g., text or image), the latest date of data acquisition or collection, language, data sources (e.g., publicly available data or privately acquired data), and the nature of the content (e.g., personal data or machine-generated data). For web-crawled data, providers must disclose information such as crawler names or identifiers, crawler behavior (e.g., whether the crawler respects robots.txt instructions), and the types of websites harvested (e.g., social media sites, cultural heritage institutions, or government portals). The most detailed requirement is “a list of

those most relevant internet domain names (top- and second-level domains, e.g., example.com) by listing the top 10% of all domain names" (European Commission 2025, p11). This disclosure requirement differs from the item-level provenance sought by some archival organizations in two important ways. First, it concerns training-data provenance rather than output provenance. In other words, it asks what data were used to train the model rather than what specific record contributed to a particular output. Second, it requires high-level information about training corpora rather than information about individual items used in training.

The United States currently has no comprehensive federal AI law requiring AI companies to disclose training-data provenance. California's Assembly Bill 2013, the Generative AI Training Data Transparency Act, requires developers of publicly available generative AI systems to publish a high-level summary of training data, including the sources and owners of datasets, whether datasets contain copyrighted or personal data, the approximate size of the datasets, whether the data were purchased or licensed, the time periods during which data were collected, and modifications made to the datasets. The law does not require disclosure of exact dataset lists, URLs, data hashes, or the complete training corpus (Chan 2026). In response to this requirement, companies such as OpenAI and Anthropic have published training-data disclosures that provide only high-level descriptions of their training sources (Anthropic 2025). For example, OpenAI (2026) states:

> "We develop these systems using a variety of data sources, including publicly available data, data that we partner with third parties to access, and information that our users or human trainers and researchers provide or generate. We also develop our systems using synthetic data."

However, even this high-level transparency requirement has not been accepted by all AI companies. For example, xAI has challenged the law and filed a lawsuit seeking to halt its enforcement (Chan 2026).

Barriers to Providing Item-Level Provenance for AI Outputs

There are significant technical, economic, and legal barriers that keep AI companies from providing item-level provenance for AI-generated outputs. First, unless a model is specifically engineered for provenance tracing, it is technically difficult or impossible, for a pretrained AI model to provide reliable item-level provenance for its outputs (Hohensinner et al. 2026; Li et al. 2026; Mo et al. 2026; Wei et al. 2026). AI models are typically trained on vast and heterogeneous collections of information and often undergo multiple rounds of training. The same fact may be learned from numerous sources. For example, a model may learn that George Washington had no biological children from books, documentaries, Wikipedia articles, archival records, or various combinations of these sources. The model may also derive this conclusion by synthesizing information from multiple sources. During training, information from these sources becomes distributed across the model's parameters rather than remaining explicitly linked to particular records. Consequently, tracing the influence of specific source items on individual AI outputs is onerous and impractical.

Second, providing detailed provenance information may create both economic and legal risks for AI companies. Information about training-data sources is often treated as a trade secret because of

its competitive value (Hohensinner et al. 2026). At the same time, AI companies face increasing legal scrutiny regarding their use of copyrighted materials. Numerous lawsuits have been filed by rightsholders alleging that copyrighted works were used without authorization or compensation (Brittain 2026; Korosec 2026). Detailed provenance disclosures could make it easier to identify evidence of copyright infringement. Studies have shown that AI models can sometimes reproduce exact or near-verbatim portions of copyrighted works contained in their training datasets (Freeman et al. 2024). In response, AI companies have implemented safeguards designed to reduce the likelihood of verbatim reproduction (Magnusson Law 2023; Stone Law 2026). These safeguards may also make it more difficult to identify direct relationships between model outputs and specific training documents. These economic and legal considerations have contributed to a broader decline in data transparency. For example, Longpre et al. (2025) found that, for the first time in 2025, downloads of open-weight models on Hugging Face exceeded downloads of truly open-source models that disclosed both model weights and training data. During the same period, the proportion of downloads associated with models that disclosed their training data declined from 79.3% in 2022 to just 39% in 2025.

Rethinking the loss of Control

Archival organizations generally have full control over collections for which they hold ownership and custody. They determine how their collections are described, preserved, and, when appropriate, deaccessioned. They may also restrict access to certain materials because of legal and ethical considerations or block service-disrupting web crawlers from accessing their online collections. Archival institutions also exercise control over the reproduction, publication, and digitization of their collections, which produce copies that leave their custody. For example, the Harvard University Archives (2026) require users to obtain permission before publishing direct quotations, excerpts, or images from their collections. NARA may decline reproduction requests for certain records because of copyright restrictions or donor-imposed limitations. In its digitization agreement with Ancestry, NARA (2015) established detailed conditions governing the resulting digital copies. During an embargo period, Ancestry was permitted to publish the digitized materials exclusively on its own website, but it was also required to provide direct links to NARA's catalog, maintain searchable databases of the digitized records, provide the ability to browse them in the order that the records appear in the boxes at NARA, and provide free access to those records at all NARA facilities. During the embargo period, NARA agreed not to sell, make available for downloading or provide access through an application programming interface. After the embargo expired, NARA will have full and unrestricted rights to use the digital copies donated by Ancestry.

Nevertheless, archival organizations generally do not expect to control the subsequent use of reproduced, published, or digitized copies once those copies leave their custody. For example, when NARA provides reproductions of records to users, it places responsibility for subsequent use on the user, including determining copyright status and obtaining any necessary permissions for copying, publication, or other uses (NARA 2023). Similarly, the Champaign County Historical Archives (2024) state that they take “no responsibility for infringement of copyright laws, invasion of privacy, or any other improper or illegal use that may arise from reproduction of any image.” Dave Hansen (2026), an attorney and librarian, argues that the remedy for misuse lies downstream in law, professional norms, and scholarly ethics rather than in restricting access in anticipation of possible misuse. For example, if someone were to use an oral-history recording of living

individuals to create deepfakes and cause harm, they should be subject to legal penalties. Archival organization should not restrict access to those collections for fear that such misuse could happen.

Nor do archival organizations expect to control the archival knowledge absorbed into human brains. When a researcher studies archival materials, the information acquired becomes part of that person's knowledge. Researchers may later share that knowledge through conversations, lectures, blog posts, articles, or books. They may combine it with knowledge from other sources or transform it into creative works such as paintings, musical compositions, or poetry. Archivists generally do not expect to control how that knowledge is subsequently used, nor do they expect individuals to forget information learned from archival materials.

If an archival organization permits some of its collections to be used for AI training, it needs only to provide a digital copy of those materials while retaining ownership and custody of the original records. AI training does not alter the legal ownership, custodianship, or physical control of archival collections. What archives may lose control over is the knowledge absorbed into AI models. Once learned by a model, that knowledge may be difficult or impossible to remove completely, and archival institutions may have limited ability to govern its downstream use. However, given that archival organizations have traditionally not expected to control the downstream use of published, reproduced, or digitized copies of their collections, nor the archival knowledge absorbed into human brains, it is worth asking why they should expect to exercise such control over knowledge absorbed into AI models.

One possible answer is that AI models can generate benefits from archival knowledge at a scale that far exceeds that of individual human users. Frontier models, such as GPT and Claude, serve millions or even billions of users. They support not only individual users but also businesses, software applications, AI agents, and numerous downstream products and services. These systems generate substantial revenue through subscriptions, licensing agreements, and related business models. Archival organizations themselves may benefit from increasingly capable AI systems, yet they typically must pay to access advanced AI services. Some archivists also worry that if archival materials become absorbed into AI models, users may satisfy their information needs through AI systems rather than consulting archival repositories directly. Reduced use of archival collections could affect institutional visibility, performance evaluation, and funding opportunities. From this perspective, under-resourced archival institutions may feel that it is unfair to contribute valuable resources to wealthy technology companies without receiving commensurate benefits in return. The concern about “loss of control” may in fact be anxiety about exploitation.

Archival organizations should carefully select collections for AI model training so that the need to remove archival knowledge from AI models does not arise in the first place. If an archival organization later changes its mind and decides to remove certain information from a model, machine-unlearning techniques may provide at least partial remedies. Machine unlearning refers to techniques that enable AI models to selectively remove or “forget” specific information learned during training. The concept has emerged largely in response to privacy concerns and legal requirements, including debates surrounding the European Union's “right to be forgotten” under the General Data Protection Regulation (GDPR) (Nguyen et al. 2025). Common machine-unlearning approaches include: (1) removing unwanted data and retraining the model from scratch;

(2) model editing; (3) fine-tuning the model away from the unwanted information; and (4) filtering unwanted outputs (Xu et al. 2024). Retraining can eliminate the influence of specific training data, but while this may be economically feasible for smaller models, it is generally cost-prohibitive for large foundation models. Model editing is also challenging because information derived from a particular source may be distributed across millions or even billions of parameters, making it difficult to identify and modify all affected parameters. The latter two approaches do not remove information from the model itself; rather, they suppress the appearance of that information in model outputs. Although none of these methods is perfect, their combined use may partially address archivists' concerns about preventing certain archival knowledge from appearing in AI-generated outputs.

Potential Costs of Refusing AI Training Requests

If archival organizations decline requests to use their collections for AI model training, they may forgo potential benefits associated with collaboration with AI companies, such as funding for digitization and access to AI services at reduced or no cost. At the same time, to fulfill their missions and meet user demands, many archival institutions would still have to digitize and publish their collections for online access. AI developers may still be able to obtain access to digitized and published online archives through web crawling, albeit at a slower pace.

To prevent AI companies from using their online collection for AI training, archival organizations could use robots.txt files to block AI crawlers. Current robot.txt provides two options to block crawlers. The first is to name each individual crawler to be blocked. However, it is difficult to identify and maintain an exhaustive list of AI-related crawlers because new developers continue to emerge and not all publicly disclose their identities. Once archival content has been absorbed into one single open-weight model, knowledge derived from that content may spread throughout the broader AI ecosystem through model reuse, distillation, weight merging, and the creation of derivative models. An alternative approach is to block all crawlers in the robots.txt file. However, this would create unintended consequences, such as preventing their online collections from being indexed by search engines such as Google, or being collected by Internet Archive and Common Crawl, which are also used for non-AI purposes, such as academic research and preservation of human culture. In addition, robots.txt remains a voluntary standard and lacks effective enforcement mechanisms. Evidence suggests that some AI developers have ignored or bypassed robots.txt opt-out instructions (Longpre et al. 2024b).

Archival organizations could use more advanced techniques to defend against aggressive AI crawlers. For example, the University of North Carolina Libraries implemented IP blocking, request throttling, regional traffic prioritization, subnet-based blocking, commercial web application firewalls, and in-browser client verification. However, the crawlers repeatedly adapted to these mitigation efforts and developed ways to circumvent many of the controls (Casden et al. 2025). In addition, these measures make it more difficult for AI crawlers to scrape online collections, but they do not make scraping impossible. More restrictive approaches, such as token-based authentication and requiring user registration and login, can provide stronger control over access and reduce the feasibility of large-scale harvesting. However, these approaches also make it difficult for legitimate users seeking access to the collections.

Archival organizations also have the option to apply website terms of service that prohibit commercial AI model training. To make it more effective, these terms can be implemented as click-through terms or API conditions (University of Edinburg 2025). In practice, however, such restrictions can be difficult to enforce. As discussed earlier, AI companies are generally neither willing nor legally required to disclose detailed training-data provenance, and many have implemented safeguards that reduce direct reproduction of source content. Consequently, even when archival materials have been used in model training, it may be difficult for archival organizations to prove that. Even if they can, the cost of going after violators might be prohibitive for many under-resourced archival organizations. The consequence could be that good actors, like academic researchers, will avoid use. And bad actors will continue undeterred (Hansen 2026).

All the measures mentioned above for blocking online archive collections from commercial AI training, regardless of their effectiveness and enforceability, entail costs and could cause unintended consequences. Refusing to directly work with AI companies or data suppliers would force them back to the unethical or illegal behaviors of scraping online resources with no consent or compensation. In addition, it could also perpetuate the bias and lack of diversity in AI training data.

Proposed strategies

AI models are unlikely to threaten the existence of archival organizations. The fact that AI companies are unable, unwilling, and not legally required to provide reliable item-level provenance for AI-generated outputs may actually constitute a competitive advantage for archival repositories. AI models trained on archival materials cannot satisfy all user needs. This limitation reinforces the continuing importance of archives as trusted sources of evidence. Users who require item-level provenance, contextual information, or access to original records will continue to rely on archival repositories. Moreover, unlike Wikipedia articles, which can be summarized, synthesized, or regenerated from existing sources, archival records are primary sources created through human activity. Their uniqueness and evidential value cannot be replicated by AI models. Although concerns that AI models may reduce traffic to archival repositories are understandable, it is equally plausible that AI systems could facilitate knowledge discovery and direct more users to archives for verification, contextualization, and further research.

I therefore suggest that archival organizations adopt a more cooperative approach toward AI-training requests. Collaboration with AI companies could provide resources that help alleviate the longstanding problem of under-resourcing in archives, reduce processing backlogs, accelerate digitization efforts, expand online access, and potentially support new acquisitions. Just as archival organizations have negotiated agreements with digitization partners, they can establish mutually beneficial arrangements with AI companies or their data suppliers. I suggest that archival organizations consider the following principles when negotiating such agreements: 1. Lower expectations. Archival organizations should not expect AI models to provide reliable item-level provenance for their outputs, to completely remove unwanted archival content from trained models, or to destroy or decommission models when complete removal is technically infeasible. 2. Preserve competitive advantages. The original archival records and the finding aids that describe contextual relationships among those records constitute key competitive advantages of archival repositories. Archives might permit AI companies to train models on primary records while excluding finding aids from training datasets. They could also prohibit the generation of outputs that reproduce

archival materials verbatim or that closely mimic original archival content. Such measures could also help address the concern raised by some moving-image archives mentioned earlier in this paper: the risk of reputational damage or financial loss if AI models are used to generate harmful content or derivative works that closely resemble original archival media. 3. Require indemnification for downstream misuse. Archival organizations should make clear that AI companies and model users are responsible for complying with applicable laws and regulations when using archival knowledge absorbed into AI models. Archives should not be held responsible for illegal or unethical downstream uses of such knowledge. A useful precedent comes from the Champaign County Historical Archives (2024), whose reproduction guidelines state:

> "In all instances, the [reproduction] applicant agrees to hold the Champaign County Historical Archives and its agents harmless against any and all claims arising or resulting from the use of this image and shall indemnify the Champaign County Historical Archives and its agents for any and all costs and damages arising or resulting from any such unauthorized use."

This principle could be extended to AI-training agreements, ensuring that responsibility for misuse rests with AI companies and model users rather than with archival institutions.

Tailor policies according to the legal status of individual collections

Archival organizations should adopt different strategies for collections with different legal statuses.

1. Collections subject to access restrictions
Collections subject to access restrictions because of privacy concerns, ethical considerations, donor agreements, or acquisition contracts should be excluded from AI training.

2. Public-domain collections
Although archival organizations do not hold copyright in public-domain materials, they retain physical custody of these collections and often have invested substantial staff time, expertise, and financial resources in their management and preservation, often over decades. Archival organizations could adopt approaches similar to those pursued by Harvard University Libraries and the Bodleian Libraries at the University of Oxford (Im and Kumar 2025; University of Oxford 2025). Under this model, AI companies provide funding for the digitization of public-domain archival materials, while the resulting digital collections are made openly available for all uses, including both commercial and non-commercial AI training. Archival organizations could also negotiate additional cost-recovery terms.

3. Collections for which archival organizations hold copyright
Different strategies may be appropriate depending on institutional missions and collection types. Archives committed to public access and open-knowledge principles, such as university archives and special collections, may choose an approach similar to that of Wikimedia. Wikimedia provides free and open access to its content for ordinary users while charging for high-volume, high-frequency commercial access through Wikimedia Enterprise (Becker 2026). By contrast, archival organizations whose collections have substantial commercial value, such as some moving-image archives, could follow the example of news organizations, scholarly publishers, and stock-image

companies by licensing access to collections for AI-training purposes (Associated Press 2023; Shutterstock 2023). For example, OpenAI has entered into licensing agreements with several major publishers, including The Associated Press, Axel Springer, and various magazine publishers (Paul and Tong 2024).

4. Collections with unclear or undetermined copyright status

For collections whose copyright status is unclear or cannot be determined, archival organizations could adopt approaches similar to those already used for reproduction and digitization activities. For example, when responding to requests to reproduce records with uncertain copyright status, NARA states that it will generally make such materials available with a warning that responsibility for subsequent use rests entirely with the user (NARA 2023). Some archival institutions digitize and provide access to orphan works or materials of uncertain copyright status after conducting a good-faith search for rights holders, often accompanied by a takedown policy inviting rights holders to contact the institution if they object to online access (SAA 2026; The National Archives 2021). When considering AI-training requests, risk-tolerant archival organizations might combine collections of uncertain copyright status with collections for which they hold copyright and negotiate compensation for their use in AI training. Should a copyright holder later emerge, the archival organization could pass that compensation to the rights holder. More risk-averse institutions may choose to treat these collections in the same way as restricted collections and exclude them from AI training altogether.

5. Collections for which copyright is owned by third parties

When copyright is owned by parties other than the archival organization, the archive generally cannot negotiate AI-training agreements on behalf of the copyright owner. However, archival organizations may assist AI companies or data suppliers in locating and contacting the relevant rights holders. For example, NARA states that it may provide users with information acquired with the material, including accession files, agency catalogs, scripts, production files, and deeds of gift. NARA also maintains a Source and Permission Contact List (https://www.archives.gov/research/motion-pictures/source-and-permission-contacts) for certain collections, directs users to the United States Copyright Office to search copyrighted works, or seek guidance of rights clearance specialists (NARA 2023). Similarly, the Champaign County Historical Archives indicate that they will assist reproduction applicants in "contacting individuals pertaining to permissions by providing addresses, when available."

6. Collections with absent or unclear donor intentions

Because AI-related uses of archival materials are relatively new, many archival organizations have not yet revised their deeds of gift to address AI training. As a result, donor intentions regarding AI use are often absent or ambiguous. In developing policies and donor agreements, some institutions have assumed that donors would oppose the use of their materials for AI model training (Lo 2026). I suggest that archival organizations conduct empirical research to better understand donor attitudes before making policies. It is possible that some donors would support the use of their materials for AI training, particularly when such use provides compensation, expands access, or contributes to broader public benefits.

Use Retrieval-Augmented Generation (RAG) technology for item-level provenance

RAG is a technique designed to enhance the capabilities of large language models (LLMs) by combining external information retrieval with language generation to produce more accurate, contextually relevant, and up-to-date outputs (Kimothi 2025). In a RAG architecture, content is indexed, stored, and retrieved from an external knowledge source rather than from model weights. As a result, reliable item-level provenance can be provided because the underlying source materials remain identifiable and retrievable. RAG technology is already being used to provide access to archival resources through its implementation in search systems, chatbots, and AI assistants. For example, when a user asks a question in Google Search AI Mode, the Gemini model is first used to analyze the user's query, then RAG technology is used to retrieve relevant information from the web, and Gemini is subsequently used to generate a synthesized response accompanied by links to source materials (Google 2026). Similarly, the Primo Research Assistant, now deployed in many libraries, is built on a RAG architecture (Ex Libris 2025). It searches the Ex Libris Central Discovery Index (CDI), which can be supplemented with local library resources, including digital libraries, institutional repositories, and online archival collections (Ex Libris 2020). Some archival institutions have also begun developing their own RAG-based systems to improve access to archival materials. One example is the Library of Congress's FolkRAG project (Kelly and Jafari 2025).

Building on existing online finding aids, digital archival collections, and aggregators such as the Digital Public Library of America (DPLA) (https://dp.la/), ArchiveGrid (https://researchworks.oclc.org/archivegrid/), and Social Networks and Archival Context (SNAC) (https://snaccooperative.org/), the archival community could develop a shared RAG infrastructure for archival discovery and access. Such a system would represent an AI-enhanced version of the National Network of Online Finding Aids proposed by the California Digital Library and its partners (Turner et al. 2024). Similar to ArchivesSpace (https://archivesspace.org/) and DPLA, this shared RAG system could be collectively developed, governed, and maintained by the archival community.

Push for government policy intervention

Even if archival organizations receive compensation through mutually beneficial agreements with AI companies, such compensation will apply only to archival content used for future AI training. By the time compensation mechanisms are established, AI companies already have trained their models on substantial amounts of online archival content. Archives are not unique in this regard; similar concerns affect many other creators and institutions whose content has been incorporated into AI training datasets. In addition, AI companies use user-interaction data automatically collected through their AI services to further train and improve their models (OpenAI 2026b). Clearly, AI models are built upon the collective intelligence of human society. However, they are negatively affecting and, in some cases, even threatening the institutions and communities that sustain parts of this collective intelligence. For example, the Wikimedia community has expressed concerns that Wikipedia could become a dark substrate for AI systems while experiencing declining traffic, reduced visibility, and potential losses of contributors and donors (Tarkowski and Valdelli 2026). More broadly, AI technologies have contributed to job displacement and layoffs across many occupations (World Economic Forum 2025). Ironically, public funds derived in part from taxes paid by human workers have been used to support the development of AI

technologies that may displace those same workers. (Yang 2026). These are valid concerns, which have led some archival organizations, including the RAC, to refrain from using AI technologies internally and to decline participation in AI training projects (RAC 2026).

Even AI companies themselves have acknowledged these problems and proposed several ways for the public to share the economic benefits generated by AI, including providing free or low-cost access to the basic capabilities of foundation models, increasing taxes on capital gains at the highest income levels, taxing automated labor, and creating a Public Wealth Fund that provides every citizen with a stake in AI-driven economic growth (OpenAI 2026c). Archival organizations should therefore join other stakeholders in advocating for public policies that ensure AI-generated wealth contributes to public goods, including cultural heritage institutions. If AI companies become more visible contributors to the public good, concerns about exploitation may diminish. In turn, archival organizations may become more willing to collaborate with AI companies under carefully negotiated arrangements that advance both technological innovation and sustainability of archives.

Conclusions

Archival organizations, as custodians of large quantities of unique cultural heritage materials that remain unpublished or undigitized, are likely to receive increasing numbers of requests from AI developers and their data suppliers seeking access to collections for AI training. Rather than adopting a presumption against collaboration, archival organizations may benefit from engaging AI companies through carefully negotiated agreements. Such agreements should preserve the competitive advantages of archival repositories, indemnify them against liability arising from the illegal or unethical downstream use of archival knowledge absorbed into AI models, and provide compensation or other benefits that help address the longstanding problem of under-resourcing in archives. At the same time, archival organizations can rely on RAG technologies to provide reliable item-level citations and advocate for government policies that encourage AI companies to contribute more to public goods and mitigate their negative impacts on society's collective intelligence, labor markets, and the natural environment.

References

Anthropic (2025) Training Data Documentation Pursuant to California Civil Code Section 3111 (AB 2013). https://trust.anthropic.com/resources. Accessed 10 Oct. 2026

Associated Press (2023) AP, OpenAI agree to share select news content and technology in new collaboration. https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/. Accessed 10 Oct. 2026

Becker L (2026) The cost of “free”: How Wikimedia Enterprise protects Wikipedia in the AI era. Wikimedia Foundation. https://wikimediafoundation.org/news/2026/07/16/wikimedia-enterprise-protecting-wikipedia-ai/. Accessed 10 Oct. 2026

Brittain B (2026) Music publisher sues Anthropic, Suno over AI training. Reuters. https://www.reuters.com/legal/legalindustry/music-publisher-sues-anthropic-suno-over-ai-training-2026-08-17/. Accessed 10 Oct. 2026

Casden J, Romani D, Shearer T, Campbell J (2025) Mitigating Aggressive Crawler Traffic in the Age of Generative AI: A Collaborative Approach from the University of North Carolina at Chapel Hill Libraries. Code4Lib J (61). https://journal.code4lib.org/articles/18489. Accessed 13 Oct. 2026

Champaign County Historical archives (2024) Reproduction use form and reproduction guidelines. Obtained through a personal contact.

Chan W (2026) AI Training-Data Disclosure Law Faces First Federal Appeals Court Test. https://www.techtimes.com/articles/320688/20260716/ai-training-data-disclosure-law-faces-first-federal-appeals-court-test.htm. Accessed 13 Oct. 2026

Data Foundation (2025) Data provenance in AI. https://datafoundation.org/news/reports/697/697-Data-Provenance-in-AI. Accessed 13 Oct. 2026

European Commission (2025). Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models required by Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act). https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models. Accessed 13 Oct. 2026

Ex Libris (2020) Central Discovery Index: A wealth of academic resources at your fingertips https://exlibrisgroup.com/products/primo-discovery-service/content-index/. Accessed 13 Oct. 2026

Ex Libris (2025) Getting Started with Primo Research Assistant.

https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_%28English%29/01. Accessed 13 Oct. 2026

Freeman J, Rippe C, Debenedetti E, Andriushchenko M (2024) Exploring memorization and copyright violation in frontier LLMs: A study of the New York Times v. OpenAI 2023 lawsuit (arXiv:2412.06370). arXiv. https://arxiv.org/abs/2412.06370. Accessed 13 Oct. 2026

Google (2026) Get AI-powered responses with AI Mode in Google Search, https://support.google.com/websearch/answer/16011537?hl=en&co=GENIE.platform%3DDesktop. Accessed 13 Oct. 2026

Tarkowski, A, Valdelli I (2026) Wikimedia, the commons, and the new AI knowledge loop: A mission for 2030, https://openfuture.eu/wp-content/uploads/2026/07/260721-Wikimedia_the_commons_and_the_new_Ai_knowledge_loop.pdf. Accessed 13 Oct. 2026

Hansen D (2026) Library and Archives 101: AI and the False Promise of Control, https://www.authorsalliance.org/2026/03/16/library-and-archives-101-ai-and-the-false-promise-of-control/. Accessed 13 Oct. 2026

Harvard University Archives (2026) Harvard University Archives Collections

https://library.harvard.edu/libraries/harvard-university-archives/collections. Accessed 13 Oct. 2026

Hohensinner R, Mutlu B, Estrada IGM, Vukovic M, Kopeinik S, Kern R (2026) Tracing the data trail: A survey of data provenance, transparency and traceability in LLMs (arXiv:2601.14311). arXiv. https://arxiv.org/abs/2601.14311. Accessed 13 Oct. 2026

Im DJ, Kumar NS (2025) How a Harvard initiative is translating archives for AI models. The Harvard Crimson. https://www.thecrimson.com/article/2025/9/17/AI-IDI-Harvard-Archives/. Accessed 13 Oct. 2026

Kimothi A (2025) A Simple Guide to Retrieval Augmented Generation (1st ed.). Manning, New York

Kelly P, Schild J, Jafari A (2025) FolkRAG: a retrieval-augmented generation system for cultural heritage materials. Neural Comput & Applic 37: 20281–20297. https://doi.org/10.1007/s00521-025-11455-4. Accessed 13 Oct. 2026

Korosec K (2026). Anthropic's landmark $1.5B copyright settlement is approved. TechCrunch. https://techcrunch.com/2026/07/20/anthropics-landmark-1-5b-copyright-settlement-is-approved/. Accessed 13 Oct. 2026

Li X, Banburski-Fahey A, Lanier J (2026) DataDignity: Training data attribution for large language models (arXiv:2605.05687). arXiv. https://arxiv.org/abs/2605.05687. Accessed 13 Oct. 2026

Lo SL (2026) The University of Virginia archival AI protocol: An AI training and access standard for archival organizations. University of Virginia. https://libraopen.library.virginia.edu/server/api/core/bitstreams/36ef8adf-ef8d-4574-88e0-2053a699d10b/content. Accessed 13 Oct. 2026

Longpre S, et al. (2025) Economies of open intelligence: Tracing power and participation in the model ecosystem (arXiv:2512.03073). arXiv. https://arxiv.org/abs/2512.03073: Accessed 13 Oct. 2026

Longpre S, et al. (2024a) Data authenticity, consent, and provenance for AI are all broken: What will it take to fix them? (ICML 2024 Spotlight). arXiv. https://arxiv.org/abs/2404.12691. Accessed 13 Oct. 2026

Longpre S, et al. (2024b) Consent in crisis: The rapid decline of the AI data commons. Advances in Neural Information Processing Systems, 37: 108042–108087.

Magnusson Law (2023) OpenAI's copyright shield: Peace of mind for some from something. https://www.magnussonlaw.com/publications/openais-copyright-shield-peace-of-mind-for-some-from-something/. Accessed 13 Oct. 2026

Mo WJ, et al. (2026) DebugLM: Learning traceable training data provenance for LLMs (arXiv:2603.17884). arXiv. https://arxiv.org/abs/2603.17884. Accessed 13 Oct. 2026

NARA (2015) Digitization agreement national archives and records administration (NARA) and ancestry.com operations inc. (Ancestry) https://www.archives.gov/files/digitization/pdf/ancestry-2015-agreement.pdf. Accessed 13 Oct. 2026

NARA (2023) Permissions/rights for use of special media records at NARA. https://www.archives.gov/research/motion-pictures/permissions. Accessed 13 Oct. 2026

Nguyen TT, et al. (2025) A survey of machine unlearning. ACM Transactions on Intelligent Systems and Technology, 16(5):1-46.

OpenAI (2026a) Training Data Summary Pursuant to California Civil Code Section 3111, https://help.openai.com/en/articles/20001044-training-data-summary-pursuant-to-california-civil-code-section-3111. Accessed 13 Oct. 2026

OpenAI (2026b) How ChatGPT learns about the world while protecting privacy, https://openai.com/index/how-chatgpt-protects-privacy/. Accessed 13 Oct. 2026

Open AI (2026c) Industrial Policy for the Intelligence Age: Ideas to Keep People First https://cdn.openai.com/pdf/561e7512-253e-424b-9734-ef4098440601/Industrial%20Policy%20for%20the%20Intelligence%20Age.pdf. Accessed 13 Oct. 2026

Paul K, Tong A (2024) Inside Big Tech's underground race to buy AI training data. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/?utm_source=chatgpt.com. Accessed 13 Oct. 2026

RAC (2026) AI Use in Our Archival Processes https://docs.rockarch.org/ai-archival-processes/. Accessed 13 Oct. 2026

SAA (2026) Issue brief: orphan works. https://www2.archivists.org/statements/issue-brief-orphan-works. Accessed 13 Oct. 2026

Shutterstock (2023) Shutterstock expands partnership with OpenAI, signs new six-year agreement to provide high-quality training data. https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year. Accessed 13 Oct. 2026

Stone Law (2026) Concord Music Group v Anthropic settles guardrails dispute. https://www.stoneslaw.net/concord-music-group-v-anthropic-settles-guardrails-dispute/. Accessed 13 Oct. 2026

The National Archives (2021) Orphan Works Guidance for archive services on reasonable searches to identify rights holders,

https://cdn.nationalarchives.gov.uk/documents/archives/orphan-works-guidance.pdf. Accessed 13 Oct. 2026

TAI (2026a) Content licensing language template: limited use, https://www.trustarchives.org/content-licensing-limited-use/. Accessed 13 Oct. 2026

TAI (2026b) Strategic engagement with tech companies: Audio visual archives engaging with AI tech companies. https://www.trustarchives.org/tools-strategic-engagement/. Accessed 13 Oct. 2026

TAI (2026c) Content licensing language.
https://www.trustarchives.org/tools-content-licensing/. Accessed 13 Oct. 2026

Turner A, et al. (2024) Building a National Finding Aid Network Final Report, https://escholarship.org/uc/item/6xr6q38m. Accessed 10 Oct. 2026

University of Edinburg (2025) Protecting University Repositories from Aggressive Web Scraping – Using Database Rights to Retain Control Over Academic Content.
https://libraryblogs.is.ed.ac.uk/blog/2025/06/12/protecting-university-repositories-from-aggressive-web-scraping-using-database-rights-to-retain-control-over-academic-content/?utmsource=chatgpt.com. Accessed 10 Oct. 2026

University of Oxford (2025) Oxford and OpenAI launch collaboration to advance research and education. https://www.bodleian.ox.ac.uk/about/media/oxford-and-openai-launch. Accessed 10 Oct. 2026

University of Virginia Library (2026) Statement of shared practice: AI training requests and unique cultural collections (Version 1.2).
https://libraopen.library.virginia.edu/server/api/core/bitstreams/66d5b442-57e6-4c60-bfe5-5a1fd5553b03/content. Accessed 10 Oct. 2026

University of Washington Libraries (2026) Statement of shared practice on AI training models and archival collections. https://library.washu.edu/news/statement-of-shared-practice-on-ai-training-models-and-archival-collections/. Accessed 10 Oct. 2026

Wei J, et al. (2026). GenProve: Learning to generate text with fine-grained provenance. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 5027–5048.

World Economic Forum (2025) Future of Jobs Report 2025: 78 Million New Job Opportunities by 2030 but Urgent Upskilling Needed to Prepare Workforces.
https://www.weforum.org/media/future-of-jobs-report-2025-78-million-new-job-opportunities-by-2030-but-urgent-upskilling-needed-to-prepare-workforces/. Accessed 10 Oct. 2026

Xu J, et al. (2024) Machine unlearning: Solutions and challenges. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(3): 2150-2168.

Yang A (2026) We’re subsidizing a technology that will replace millions: Andrew Yang on calling for AI tax. https://www.cnbc.com/video/2026/08/18/were-subsidizing-a-technology-that-will-replace-millions-andrew-yang-on-calling-for-ai-tax.html. Accessed 10 Oct. 2026