Invest in Open Infrastructure

Sustaining the Commons in the AI Economy

A Landscape Scan of Challenges and Strategies for Bridging AI Companies and Open Curated Collections

Sarah Lippincott
Invest in Open Infrastructure
https://orcid.org/0000-0002-5700-5844

Lauren Collister
Invest in Open Infrastructure
https://orcid.org/0000-0001-5767-8486

Katherine Skinner
Invest in Open Infrastructure
https://orcid.org/0000-0003-0139-7524

21 April 2026

Executive Summary

This landscape scan was conducted by Invest in Open Infrastructure in the context of the Building Resilient Infrastructure through Dialogue, Growth, and Exchange project exploring how stakeholders in the knowledge ecosystem can work together as AI reshapes how collections are used and valued.

Curated collections form part of the digital commons: shared open resources maintained academic institutions, non-profit organizations, governments, and private companies. Sustained largely by public and private funding and an ethos of open knowledge sharing, these collections represent a public good inseparable from the health of open science and democratic access to information. **That infrastructure is now under strain.** AI labs bring insatiable demand for immediate access to high-quality training data. Automated bots now generate traffic that, in some cases, exceeds human visits, overwhelming servers and inflating costs. AI-generated submissions threatens to overwhelm editorial workflows. These bots include large technology companies and a significant long tail of start-ups and individual experimenters.

In response, collections stewards have deployed bot-blocking tools, updated access policies, and pursued licensing strategies. Yet these responses have proven inadequate. Technical countermeasures are routinely circumvented. Licensing frameworks are nascent and poorly enforced. Legal protections are fragmented and oriented toward intellectual property rather than broader harms to the commons. Defensive access restrictions risk accelerating data consolidation, undermining the openness they were designed to protect.

The report points toward a more promising path: commons-based governance grounded in reciprocal norms and shared interests. Data users have concrete reasons to want the commons to survive: loss of key data sources reduces training data quality; an increasingly walled-off web raises legal risks; and public frustration creates reputational pressure. Well-maintained curated collections offer unique, authoritative content that produces better models. Investment in the commons, properly framed, is investment in the quality of AI.

Yet a central tension remains unresolved. Existing frameworks rely on voluntary compliance, while evidence shows that voluntary frameworks have so far failed to change behaviour at scale. Whether enlightened self-interest will prove more effective than the legal and technical mechanisms that have already fallen short remains an open question. Answering it requires engaging curators and consumers of open collections as stewards of the digital commons, co-creating partnership models that align open knowledge strategies with commercial demand.

Preamble

This landscape scan was conducted by Invest in Open Infrastructure (IOI) in the context of the Building Resilient Infrastructure through Dialogue, Growth, and Exchange (BRIDGE) project (supported by the Mellon Foundation, 2026–28) which explores how different stakeholders in the knowledge ecosystem can work together as developments in machine learning and AI reshape how knowledge collections are used and valued.

Through this report and through the BRIDGE project, we seek to reveal the limitations of our dominant narratives about AI and encourage new ways to work across diverse stakeholders to resolve problems that may only be able to be addressed through aligning diverse voices, drivers, and approaches. The project stakeholders — including stewards of open research and cultural heritage collections (referred to as "open curated collections" in this report), scientific publishers, commercial entities utilizing content as data for their innovation pipelines, and organisations harvesting content at web scale for LLM training and inference — will explore and design partnerships and business models that advance mutual interests, including the resilience of open knowledge collections.

This landscape scan surfaces the challenges and drivers of the AI economy; it does not offer specific solutions. Rather, it closely examines the major factors that are defining different stakeholder experiences of today's rapid transformation in the ways we encounter, use, validate, and maintain access to knowledge. We consider such an examination to be a precondition for engagement across the currently fraught "us" and "them" boundaries that currently divide different constituents in the ecosystem. The next phase of our work will build on this scan to engage stakeholders from across the ecosystem to co-create partnership models that align open knowledge strategies, revenue strategies, and commercial demand, using this report as a shared starting point. For additional information and materials from the study, visit our website.

Introduction

Across the web, a seismic shift is underway. The visitors browsing the carefully curated, open collections of digital research and cultural heritage repositories are increasingly machines. In some cases, bots now outnumber people; their automated queries consume bandwidth and strain infrastructure built for a different era of access.

As artificial intelligence (AI) systems grow more powerful and more data-hungry, these collections are becoming an essential, but largely uncompensated, input to commercial AI development. Content licensing deals are emerging with a variety of prominent publishers and media organizations, and some open curated collections, but clarity on pathways to approach these sorts of negotiations is still murky to many.

This landscape review examines the growing tension between data extraction and sustainability in the digital commons. We begin by tracing the technical, financial, and ethical pressures on open curated collections, which are largely created and maintained through government and philanthropic funding and academic subsidies. We also describe the pace and relevance challenges that AI labs and tools face in the largely unregulated commercial marketplace, which is backed by investors expecting rapid economic gains. We then survey the responses collections stewards have implemented, or are actively developing, and probe at the limits of those strategies in an environment driven by pressure on well-capitalized technology firms to innovate and compete at the highest possible speed. Our analysis turns next to the evolving data marketplace, identifying what AI companies consider “high-quality” data and how current licensing deals are structured, in order to assess whether similar models could realistically apply to open curated collections.

Perceptions of AI, machine-learning technologies, and knowledge usage and generation today tend to be both strong and divisive, with arguments fueled by concerns that largely depend on specific stakeholder experiences. For some, AI holds the promise of positive societal and economic transformations. From this vantage point, AI is a mechanism that can democratize learning, significantly narrowing the chasm between people and the accumulated knowledge of civilization. It can also amplify the speed of problem solving, and it can bring highly specialized information and interpretation tools to all people, regardless of their location or training. For others, AI is a dangerously unregulated and speculation-fueled marketplace that is being built with a free-rider mentality. It enables entrepreneurial technologists to absorb generations of curatorial work into systems that render the underlying data invisible. The resulting devaluation of data may ultimately undermine the work of data creation, curation, and maintenance, dampening the availability of new, rigorously vetted data for both individual researchers and large language model

(LLM) use. From this angle, creators and curators of information are in danger of losing the value associated with their collections while simultaneously paying high costs associated with unrelenting bot traffic. Users likewise stand to lose open access pathways to collections, including diverse users' autonomy to interpret and use these collections directly.

What is striking is that neither of these narratives is "wrong"; the intensity with which both are held tells us something important not really about AI at all, but about our experience of rapid technical developments that are altering current and future knowledge ecosystems in ways we cannot fully predict. These stories — the utopian and the extractive — are indicators of social systems trying to metabolize changes in how and by whom knowledge is produced, owned, and shared. They match the arguments we are having, in public, in courtrooms, and in comment sections, about what knowledge is for, who pays for it, who gets to benefit from it, and what obligations come with building systems that depend on generations of intellectual labour.

AI has become the screen onto which we project our oldest anxieties about access and enclosure, about the commons and the commodity — and the fierceness of the debate is less a measure of AI's novelty than of how much we already understood was at stake in those questions long before the first model was trained.

The report identifies areas of alignment and shared interests between AI companies and open curated collections, which might be codified in ways that enable collaboration and trust building with the right legal frameworks and agreements. It also explores governance-oriented approaches that move beyond individual monetization toward shared responsibility — approaches grounded in reciprocity, collective action, and long-term stewardship of the digital commons.

Data races, displacement, and the double-edge of AI

In the context of digital research infrastructure, open **curated collections** refer to purposefully assembled, organized, and maintained sets of digital resources that support research, teaching, or scholarly inquiry and are made available for free and with licenses that explicitly permit broad use and reuse. Examples include digitized or born-digital archives of cultural heritage materials, open access journals, scientific data repositories, preprint servers, and knowledge graphs, among other resources. Curated collections form part of the digital commons, the shared pool of open resources including data, cultural materials, software, and more, maintained and made freely accessible online by myriad academic institutions, nonprofit organizations, governments, and private companies around the world (Dulong de Rosnay & Stalder, 2020). These resources are generally oriented toward use and reuse rather than commercial exchange.

These collections provide large, diverse, authoritative, well-structured, and easily accessible datasets[1] that have become attractive to researchers and AI developers as rich sources of training data. While open curated collections may benefit from this increased interest, and also from the potential of AI to enhance internal workflows, AI also poses a number of existential and ethical challenges that have prompted discussion, debate, and a range of strategic responses.

AI-driven tools offer a range of potential benefits to curated collections, from automating labour-intensive tasks like metadata creation and entity linking, improving discovery and access (Cohen, 2024), enabling new forms of analysis, and supporting more inclusive, user-centred curation. As examples, AI has achieved unprecedented advances in machine transcription for handwritten manuscripts (BnF, n.d.); enabled novel approaches to discovery and retrieval of digital content (Davis, 2025); expanded multi-lingual search (Matas, 2025); and made review and curation workflows more efficient (bioRxiv, 2025).

While these are significant benefits, the insatiable demand for data from the industry has created strain on curated collections. Open curated collections now face unprecedented

[1] Datasets hereafter refers to the resources provided by open curated collections, which include, but are not limited to, data, text, images, audiovisual content, metadata, and code.

volumes of API calls, web scraping, and bulk downloads from automated systems or "bots." Stewards of curated collections increasingly "worry that swarms of AI training data bots will create an environment of unsustainably escalating costs for providing online access to collections" (Weinberg, 2025). At the same time, chat bots and agents divert human traffic as information-seekers increasingly access content through AI-driven intermediaries, in many cases never clicking through to original content sources (Delaney, 2025). **This fundamental shift in how people engage with information poses ethical and operational challenges for curated collections.** When chatbots and other AI tools selectively present or summarize information, they may lose essential contextual information (e.g., retraction notices, offensive content flags, or provenance). Disintermediation will also require new approaches to understanding user needs; revised metrics for demonstrating value (as funders and supporters often look to direct traffic and similar metrics as proxies for adoption and significance); and innovative ways to draw attention to added-value services, contribution opportunities, and other revenue-generating services typically described on the collection's website.

Various sources confirm that excessive bot traffic targeting open collections is a widespread phenomenon that is causing service disruptions and inflating costs. A 2025 COAR survey of open access repositories found that 80% of respondents had experienced service disruptions as a result of aggressive bots and crawlers, with impacts ranging "from regular service slow downs, to short downtimes, to major service outages sometimes lasting for several days" (Shearer & Walk, 2025). A post on the Lyrasis Wiki in 2024 outlined common patterns across such incidents: harvesters making millions of requests per day, often from hundreds of simultaneous IP addresses; disregard for robots.txt restrictions; disguised or constantly changing User-Agent strings that make blocking difficult; and traffic patterns that are difficult to distinguish from benign use[2] (Prater et al., 2024).

Excessive traffic can overwhelm servers, in severe cases leading to website outages. It can also generate significant expenses. According to Wikimedia, bot traffic has become some of its most expensive traffic to serve due not only to its volume but because bots visit less frequently consulted pages, leading to increased traffic to Wikimedia's core datacenter rather than the more accessible content caches in a user's region (Mueller et al., 2025). Bot traffic reportedly exceeded human traffic to some curated collections in 2025 (Mulvany, 2025). Princeton University identified 63% of its library catalog requests during a recent period as originating from bots (peaking with a swarm at 79%) (Griffith & Metz, 2025). In the open source software world, some "projects now see as much as 97 percent of their traffic originating from AI companies' bots" (Edwards, 2025). In a blog post, the open source software documentation site Read the Docs reported a crawler downloading 73 TB

[2] Benign use refers to access that abides by a collection's access and use policies and is not malicious, abusive, or reckless. Much of the community discourse refers to this as "legitimate use."

of data over the course of a single month (and around 10 TB in one day), resulting in over $5,000 in excess bandwidth charges, a staggering and unbudgeted amount for most small open source projects (Holscher, 2024).

Though widespread, the problem has affected the sector unevenly. Some collections stewards have described these challenges as manageable, while others have expressed concern that the unrelenting and increasingly sophisticated bot traffic issue will require more than stopgap technical solutions in the long run. The full scale of the problem can be difficult to measure because analytics services are often configured to screen out bot traffic, and these services use imperfect mechanisms to do so. This makes it difficult for open collections to accurately count human users and demonstrate impact through traditional usage metrics — even as the relationship between AI access and downstream impact remains poorly understood.

Despite reports that the majority of publicly available web data has already been scraped and incorporated into major corpora such as Common Crawl, signs point to traffic continuing to increase, especially as "architectures such as retrieval augmented generation enable models to draw on fresh data in response to user queries, instead of relying solely on the data they were exposed to during training" (Hardinges, 2025). Retrieval augmented generation (RAG) requires continual access to data sources, rather than one-time or periodic harvests. New AI start-ups and one-off projects have proliferated as well, enabled in part by AI itself, which puts the coding powers necessary to scrape data and build new models, tools, and services into many more hands.

In addition to ballooning bandwidth costs, curated collections stewards incur a range of indirect costs related to scraper bots, including the human labour required to mitigate disruptions, rearchitect sites to be more resilient to bot attacks, develop and implement new policies, and manually block particularly aggressive harvesters, among other jobs.

The disruption has not been limited to scraping and harvesting. Curated collections that accept user-generated content have also experienced a dramatic rise in AI-generated submissions, overwhelming editors and readers with high volumes of low-quality data (Lin et al., 2025). For example, monthly submissions to arXiv hit nearly 28,000 by late 2025, signalling a shift from linear to super-exponential growth. The timing corresponds directly to the rise of generative AI. In response, arXiv changed its review practice for computer science preprints "due to the unmanageable influx of review articles and position papers" (Boboris, 2025). The severity of the situation, if not contained, could undermine the viability of the publish, review, curate (PRC) model entirely (Greaves, 2025).

Simultaneously, research suggests LLM adoption correlates with reduced human contributions to public sharing platforms (del Rio-Chanona et al., 2024), caused by

diminishing production of original content and "fears of labour replacement or lack of attribution" if their work is used for LLM training (Huang & Siddharth, 2023). Eve (2025) describes the prospect of scholars erasing decades of progress in open access adoption, because they feel safer putting their work behind a paywall, effectively "cutting off their human readership nose to spite the AI training face."

Wariness of open licenses and public sharing stems from a decades-long history of unethical and even illegal use of open data by technology companies (Tarkowski & Warso, 2024). Numerous controversies have damaged public trust and demonstrated that some companies will disregard personal privacy, intellectual property, and other concerns in the name of innovation and in search of profit.[3] The tendency of AI tools to plagiarize and regurgitate content means that training data (including personally identifying images, copyrighted content, and sensitive information) may not only underlie the creation of new materials but also risk being fully replicated as a generative AI output without the creator's knowledge or consent (Paul & Tong, 2024). Many curated collections stewards are justifiably wary of potential misuse of their own collections and the responsibilities they have to content creators. Consent or preference signalling frameworks are nascent and may be difficult or impossible to implement both retroactively and at scale.[4]

Curated collections stewards also worry about the broader erosion of public trust in the information commons as low-quality or mis/disinformation proliferates and people increasingly turn to chat bots for quick answers rather than directly consulting information sources (Shearer, 2025).

[3] The MegaFace incident seems to have been the bellwether for concerns about (mis)use of public content for machine learning; it was notable because of privacy concerns related to the use of personally identifiable information without explicit consent for that use case (Van Noorden, 2020).
[4] Some of these frameworks are discussed in more detail in a subsequent section of this report.

When defending collections threatens openness

How open curated collections respond to AI harvesting matters. The risk of an overcorrection is as real as the risk of exploitation. As large-scale harvesting of open data for AI model training and RAG has generated negative externalities — unpriced costs such as revenue displacement, infrastructure burdens, the degradation of the digital commons, individual privacy risks — for open curated collections stewards (Chan et al., 2023), restricting access has become an increasingly common and accepted mechanism for open collections stewards to protect their infrastructure and push back against practices they view as exploitative.

Collections stewards describe existing mechanisms to identify and block unwelcome access as partially effective, but inadequate and tenuous. Blocking crawlers is a complicated process. While bots from larger, more established companies may identify themselves, there is a massive long tail of bots that do not identify themselves as such and are therefore harder to exclude (Miller, 2025). Overzealous or automated blocking processes may deny access to humans (in some cases via AI intermediaries) as well as search engine and web archiving crawlers, and can disrupt internal technical processes. A 2025 COAR survey found that repositories considered their mitigation strategies to be somewhat successful, but not entirely sufficient, and acknowledged they knew they were blocking some benign users (Shearer & Walk, 2025). Collections stewards have expressed concerns that any measures they put in place may quickly become obsolete as harvesting practices evolve.

Not all bots are related to AI, nor are all bots behaving badly. Bots serve a myriad of players, including small start-ups, independent researchers, and large corporate entities. They are designed for different functions, including vulnerability scanners, content harvesters, and AI agents. Some bots are harmful due to malicious, irresponsible, or erratic and uncontrolled behaviour. Others are well-designed and intended to be low-impact on the sites they query. By contrast, most bot-blocking is indiscriminate, affecting AI agents that perform legitimate query tasks, users from entire geographic regions, and bots that are behaving badly alike. Investing in industry standards around bot identification and verification could provide ways to encourage taxonomy-based understanding and

behavioural validation, pinpointing types of actors to improve usage analytics and to respond with targeted blocking where needed (Lloyd, 2026).

Other approaches "focus on preference signals that rely on voluntary compliance by the crawling party", such as the ubiquitous robots.txt file, open licenses, or newer tools and frameworks such as CC Signals[5] and the IETF AI Preferences standard[6] (Baack et al., 2025). AI harvesters may circumvent or ignore mitigation strategies like robots.txt and firewalls. Longpre et al., (2024) audited 14,000 web domains, identifying "a rapid crescendo of data restrictions" that, if respected, threaten to bias "the diversity, freshness, and scaling laws for general-purpose AI systems." They estimate that data sources accounting for a full quarter of the tokens comprising several major AI models (C4, RefinedWeb, and Dolma) now expressly restrict harvesting through preferences articulated in their robots.txt files.

A number of frameworks and protocols have emerged to support attribution and express creator preferences around reuse. For example, Really Simple Licensing (RSL), formalized as an industry standard in December 2025, provides an immediate technical infrastructure that collections stewards can implement to express machine-readable terms of use. Building on existing technologies like robots.txt and RSS, RSL enables content providers to specify licensing terms through XML-based declarations integrated with lightweight and accessible implementations like robots.txt files, HTTP headers, RSS feeds, and HTML elements (RSL Internet Collective, 2025a, 2025b; Viana, 2025; Ponsford, 2025). Creative Commons' "CC Signals" framework enables creators to specify contextual preferences — such as allowing educational AI use while restricting commercial applications — recognizing that "blanket opt-in and opt-out do not capture critical nuances of consent preferences, as context of use and user intentions may matter more than the act itself" (Ivanova & Ding, 2025). A range of other new licenses, such as Nwulite Obodo License, Kaitiakitanga Licenses, the Montreal License, the OpenRAIL Licenses, the Open Data Commons License, and AI2Impact Licenses aim to facilitate various aspects of data openness from community data sovereignty to limits on use by Big Tech and reintroduce some "friction" to the process of data reuse (Chandrasekhar, 2025). Chandrasekhar describes friction as a potentially necessary and advantageous condition of open data stewardship in the age of AI.

There is strong evidence that AI companies routinely ignore the terms of open licenses and preference signals, in some cases because they are not machine-readable, in others because these companies rely on Fair Use exceptions to justify their activities, and in still other cases because they see little legal risk. The 2025 GLAM-E Lab survey indicated no difference in bot harvesting between openly licensed and "merely digitally available"

[5] https://creativecommons.org/ai-and-the-commons/cc-signals/
[6] https://datatracker.ietf.org/wg/aipref/about/

collections, indicating that collections are targeted equally regardless of licensing status (Weingberg, 2025). Creative Commons itself has argued that its licenses "are not well-designed for imposing reciprocal terms on AI developers" (Hardinges et al., 2025).

Organizations that harvest open content may also see open licenses as a sign that content creators or stewards have internalized costs associated with reuse. However, these licenses were never intended to cover the current volume of access, nor the potential for a systematic undermining of the value of human creativity. The idea that stewards have internalized costs also assumes a deliberate economic trade-off, ignoring the fact that many openly licensed works are shared freely as contributions to the commons without any expectation of commercial downstream value. Stewards of open content are not forgoing revenue so much as participating in a non-commercial ecosystem that AI companies are now monetizing unilaterally.

The use of open collections for commercial products has a long history. The scale of scraping for LLMs at this moment is making the dependence of commercial products on open content more visible and more contested, and highlighting opportunities for curated collections to evolve their business models.

Fundamentally, AI companies are competing in a high-stakes market that is hard to break into and easy to lose traction in. This dynamic encourages breakneck speed over care and consideration. A company that pauses to assess the impact of its data harvesting practices risks falling behind rivals who do not. In the absence of clear regulatory guardrails, restraint is punished and speed is rewarded, and the open collections these products rely on become a shared resource that every actor has reason to exploit but none has structural incentive to protect.

The legal landscape remains fragmented across jurisdictions, with unclear frameworks ill-suited to address commons-based data and ethical AI concerns. This uncertainty has, in some cases, spurred content creators to lock down resources and pursue litigation.

The result is a double bind for curated collections stewards: efforts to protect collections may simultaneously undermine Internet openness, exclude preservation activities, and consolidate power among tech giants, all while failing to effectively prevent unauthorized AI training.

Locking down resources, monetizing data, applying stricter licenses — in these responses, some see existential threats not only to curated collections, but to the open Internet itself. Various analyses warn about the emergence of a "techno-economic oligopoly" or "digital feudalism," that would further consolidate data resources into the hands of a few massive corporations, excluding civil society and smaller commercial organizations alike (Kretschmer et al., 2024 cited in Westenberger & Farmaki, 2025; Mazzucato & Gernone, 2025; Verhulst, 2025; Wiggers, 2024a). They describe "a world where Google, Microsoft, and Meta get special access through billion-dollar licensing deals while everyone else — researchers, journalists, small businesses, individual users — gets locked out" (Hellman, 2025) or turned into "serfs" who only get access in return for sacrificing their own labor or personal data. This is not an argument against commercialisation. Rather, it is an argument for commercialisation on terms broad enough to include small AI developers, academic publishers and research institutions, not just those with the legal and financial resources to negotiate bespoke deals. When the lawfulness of scraping remains uncertain, organizations with substantial legal resources gain competitive advantage: they can absorb litigation risks that smaller entities cannot.

Other risks of this environment, where organizations look to silo, protect, and (in some cases) monetize their data, include the exclusion of crawlers that support discovery and preservation. Masnick (2025b) cites the example of Reddit blocking archiving crawls from Internet Archive due to concerns that they serve as a backdoor for AI companies to access Reddit's corpus for free, undermining its plans to turn it into a revenue stream. Masnick problematizes this approach, warning that "rather than finding ways to capture that value while preserving archival access, they're choosing to break historical preservation entirely."

Existing legal and regulatory frameworks are ill-suited to address the challenges facing open curated collections stewards. Globally, policy, legislation, and lawsuits have centred on issues of copyright infringement and personal data privacy, not on other forms of "unethical, harmful, or irresponsible use of ... content for AI training" (Tarkowski & Warso, 2023) or the use of public domain or openly licensed data. Governments have largely taken a pro-AI stance in the name of enabling economic growth (Terras, 2025).

The legal frameworks relevant to AI training and delivery vary by jurisdiction. Underlying concepts of intellectual property differ. Copyright, prevalent in the US and the United Kingdom, emphasizes economic rights, while droit d'auteur, adopted in countries like France, Germany, and Brazil prioritizes moral rights of authors to their creative works.

AI training may fall under text and data mining (TDM) exceptions; Japan has one of the broadest text and data mining (TDM) exceptions globally (Pasetti et al., 2025), joined by Israel and Singapore (Tiedrich, 2025). Europe also has TDM exceptions, while the United States relies primarily on the doctrine of Fair Use (Courtney, 2024). More restrictive laws include China's comprehensive regulations for foundation model developers, covering responsibilities such as "AI governance, training data requirements, tagging and labelling standards, data protection protocols, and safeguarding user rights" (Wu, 2023).

As of August 2025, provisions of the EU AI Act went into effect, requiring "general-purpose AI systems to comply with EU copyright laws and transparency requirements, including sharing information about training data" (Tiedrich, 2025). However, it does not resolve the "underlying question of legality" of scraping copyrighted materials (Pasetti et al., 2025). Tarkowski, & Warso (2024) assert that EU copyright law allows commercial AI training on "lawfully accessible copyrighted works," without explicit permission from rights holders "unless the rights holders have made a (machine-readable) rights reservation". The effects are unclear so far, as general-purpose AI providers have until 2028 to ensure compliance.

An important distinction remains between the use of copyrighted works as training data and the copyright status of the outputs of generative AI, which remain an edge case as a "non-expressive use" (Sag, 2023). Additionally, the nature of downstream outputs could break the upstream fair use argument supporting AI training.

High-profile investigations and lawsuits have begun to shift the risk landscape for copyrighted content. Companies are starting to hedge their legal bets, secure their data supply chains, and change behaviour in response to major lawsuits (Paul & Tong, 2024). According to some analyses, the Anthropic settlement[7] in particular underscored the substantial risks of using "pirated or unauthorized materials" in training data and should prompt more AI developers to invest in "rigorous data governance," including detailed recordkeeping regarding data provenance (Loring & Rayner, 2025; Penti & Schaal, 2025). This watershed moment may prompt more expectations on the part of content creators and providers to derive a share of the value from the reuse of their work by AI companies. As of September 2025, there were as many as 40 other pending lawsuits against AI companies

[7] This settlement refers to a $1.5 billion agreement reached as the result of a class-action lawsuit accusing the AI company Anthropic of using pirated books to train its models.

specifically regarding the use of unlicensed data (Brandom, 2025). The increasing prevalence of RAG augments the potential for copyright infringement as it tends, more than other models, to engage in direct quotations from source materials, which may undermine Fair Use arguments (Sag, 2025).

Lawsuits that aim to strengthen copyright protections[8] have been opposed by organizations such as the Authors' Alliance on the basis that they would "grant unlimited power to private parties to restrict speech" and give a handful of massive corporations "unilateral control" over how content on their platforms can be used, suppressing the rights of individual creators and potential reusers (Xu, 2025).

Legal ambiguity around the lawfulness of data harvesting "not only fuels litigation risks but also fosters defensive practices by AI developers, including opacity regarding dataset composition and reluctance to disclose sources, which undermines broader efforts toward transparency and responsible AI governance" (Pasetti et al., 2025). However, some organizations are adopting an alternative approach: training exclusively on openly licensed content (e.g., EleutherAI's Common Pile) (Biderman et al., 2025).

We need better mechanisms to sustain the commons of open data.

To develop these mechanisms, we need to understand the technical and economic value of data for AI.

[8] For example Ziff Davis v. OpenAI.

Exploring the data marketplace

Massive quantities of data are needed to build "foundation models," the generalized machine learning systems trained to perform a wide array of tasks. A small number of US-based players, including OpenAI, Anthropic, Google DeepMind, Meta, and several large Chinese companies, including Alibaba Group Holding, DeepSeek, and Moonshot AI, dominate foundation model development. Model builders may build their own datasets through scraping the web and harvesting via APIs; they may also work with data aggregators and brokers that compile and curate datasets for sale or open distribution and build their own filtered versions of large corpora (e.g., Common Crawl). These pre-trained LLMs are not "products" themselves, but "infrastructure" that underlies myriad specialized AI tools that serve a range of domains, from code generation to scholarly research, and more (Mozilla Foundation, 2024).

AI requires data for model training and inference, the process of AI-driven systems responding to a prompt.

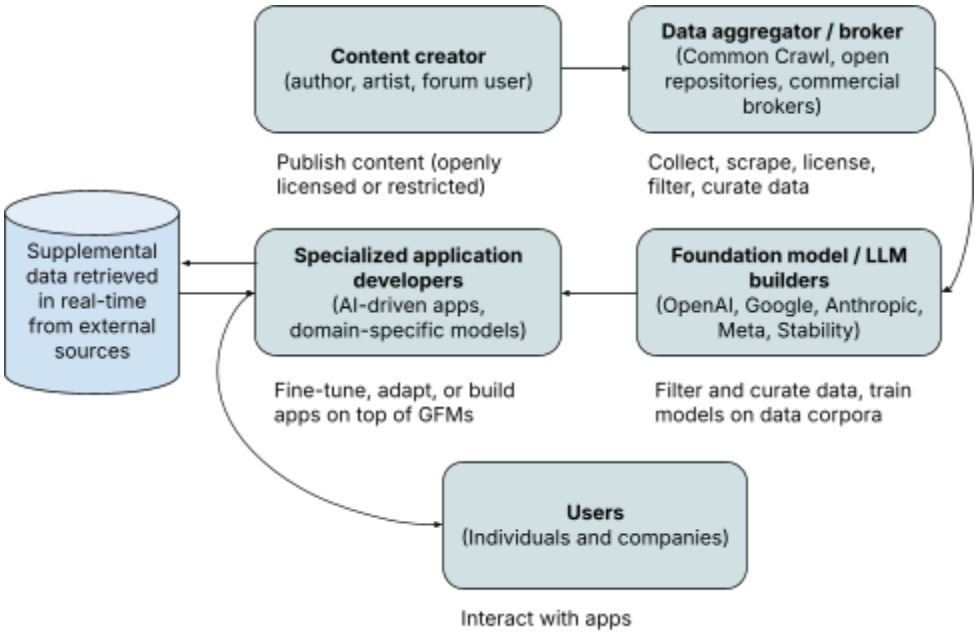

Figure 1. Data value chain from content creators to end users.

Data, not architecture, appears to account for most improvement in LLMs (Kaplan et al., 2020). Scale is important; generative AI systems use probabilistic models and therefore need vast amounts of data to effectively predict the outputs that make the most sense. Generative AI in particular needs high volumes of unstructured data such as text, images, and video. High-quality data[9] matters more than "sheer quantity" in order for models to produce high-quality outputs (i.e., relevant, accurate, convincing, safe, context-appropriate content) (Wiggers, 2024a). This has become increasingly true as AI developers move towards inference or retrieval augmented generation (RAG), which queries data in real time rather than relying solely on training data. Models trained on mixed-quality web data also need high-quality data for refinement. Hellman (2025) explains, "Good training data has lots of text, lots of metadata, is reliable and unbiased. It's unsullied by Search Engine Optimization (SEO) practitioners. It doesn't constantly interrupt the narrative flow to try to get you to buy stuff. It's multilingual, subject specific, and written by experts. In other words, it's like a library." Internal messages revealed as part of a class action lawsuit against Anthropic stated that "training AI models on books could teach them 'how to write well' instead of mimicking 'low quality internet speak'" according to reporting in the Washington Post (Schaffer et al., 2026). Wang & Jia (2023) argue that data value primarily equates to "how much a datum or dataset improves model performance" rather than its "market-price or creator-payments by default." LLM developers can use standard statistical measures to identify which samples or datasets increase accuracy, reduce loss, or improve the downstream capabilities of their models.

Current deals and data monetization

Recent legal settlements, licensing deals, and market reactions have reinforced the economic value of copyrighted content and have led some organizations to recalibrate their content harvesting methods. Companies also recognize that there are some things brute force scraping cannot provide: reliable access over time, well-documented provenance, clear usage rights, and well-structured data. The need for high-quality data, as well as an increasingly risky legal environment, explains the surge in partnerships with news organizations, academic publishers, legal archives, and code repositories.

[9] There is no single definition of "high-quality data" for machine learning. It may mean well-structured, peer reviewed, free from offensive or biased content, etc. See **Table 1** for an inventory of characteristics of high-quality data gleaned from the literature.

Characteristic	Notes
Real-time or recent	Up-to-date information helps keep models relevant and accurate. Models trained on outdated information provide obsolete answers about evolving topics, from current events to scientific discoveries to changing social norms and technical standards.
Openly licensed	Machine-readable open licenses help AI companies mitigate legal risk involved in harvesting and using content.
Authoritative and accurate	Peer-reviewed, expert-authored content helps deliver reliable information and results to end users of chatbots and generative AI tools.
Representative and diverse	Data that represents the full diversity of real-world scenarios the model will encounter helps it deliver more appropriate and useful outputs. Smaller open data providers can help diversify the data corpora used to develop foundation models.
Unique	Content not duplicated elsewhere or that represents novel perspectives provides models with distinctive patterns to learn from. This is particularly valuable for specialized domains or underrepresented topics and helps model developers differentiate their products.
Knowledge-dense and voluminous (long-form)	Knowledge-dense sources like academic papers, technical documentation, and scholarly books contain concentrated information with a high signal-to-noise ratio (i.e., proportion of useful versus irrelevant or incidental information). Long-form content helps models understand context, narrative structure, and sustained argumentation. Voluminous in this context refers to the length of a specific item, not the size of a full corpus.
Multilingual	Diverse language training data leads to more accurate and reliable performance for users who speak or want to generate content in a range of languages.
Well-structured and described	Metadata provides crucial context that streamlines learning. Models converge faster and achieve higher accuracy by leveraging consistent, well-defined input signals rather than repeatedly learning fundamental patterns.

Table 1. Characteristics of “high-quality” training data for AI models.

The price tags for these deals vary dramatically, from single-digit millions (for smaller publishers or one-off academic collections) to tens or hundreds of millions for large, multiyear news and image-library deals. In 2024, *Business Research Insights* estimated the AI-training data market at around $2.5 billion, with anticipated growth to $30 billion within a decade (Paul & Tong, 2024). More conservative analyses argue that current deals represent investor-subsidized experiments rather than an indication of a sustainable marketplace for data (Woahn, 2026). There is also evidence that the demand for training data may have already peaked, while data for RAG is growing (Vigliarolo, 2026).

Many of the largest deals over the past few years have been struck with news and media organizations. *TechCrunch* estimates OpenAI is spending between $4 million and $20 million a year for **news content** in total. They base their estimate on reporting from *The Information* that revealed OpenAI was "offering publishers between $1 million and $5 million a year to access archives to train its GenAI models," though marquee publishers command much larger deals (Wiggers, 2024b).

Such deals typically combine an upfront fee with recurring or usage-based payments (and sometimes non-cash credits, tech cooperation, or product integrations). Media and publisher deals tend to include attribution requirements when content is surfaced via the AI product, non-exclusive rights for training, and no verbatim reproduction of paywalled content. News publishers especially want to avoid "AI clones" of their content that would cannibalize their business. Specific terms and dollar amounts are largely undisclosed; a few examples of publicly available deal amounts are described in **Appendix A**.

We can derive several pricing clusters from these and other examples:

- **High-value content** (e.g., premium news archives and large social platforms with high-traffic and unique data) commands up to and over $100M, often covering a multi-year arrangement. The highest annual payment surfaced for this report was Reddit's estimated $60M per annum deal with Google.
- **Mid-tier content** (e.g., stock photo and video libraries and large academic publishers) has attracted deals ranging from $10M to $50M per year.
- **Low-value content** (e.g., user-generated content from business review platforms and Q&A forums) is netting mid-seven figures or revenue-sharing deals.

Data with clear rights, high editorial standards, large volume, or recency value appears to command higher prices. AI companies also seem to be increasingly interested in academic content because it is highly structured, reviewed, domain-specific, and expensive to produce (and therefore scarce). These factors are all crucial to the development of specialized LLMs and for refinement of general-purpose models. These data valuations are

consistent with the characteristics of "high-quality" data described in the literature and summarized in **Table 1**.

We can also deduce data value from class actions and settlements (e.g., cases involving book authors and Anthropic and other model builders) that have disclosed proposed compensation in the low thousands of dollars per title for qualifying works, though terms and final amounts vary by case (Malhotra, 2025).

Some have argued that generative models are hitting a "web-data ceiling," making it increasingly difficult to source novel, high-quality data. Some analyses predict that the most promising future opportunities for monetization therefore will lie in untapped data reserves, including sensitive, private, analog, or internal data. The non-profit organization OpenMined estimates that AI has only been "trained, evaluated and deployed on less than .01% of all data," specifically publicly available digital data scraped from the web. They contend that a further 180 zettabytes of private, sensitive, analog, and otherwise inaccessible data, presents a new frontier for AI training (Hardinges, 2025). Significant interest has also turned toward synthetic data as a compelling potential avenue for augmenting traditional human-generated corpora during the foundational pre-training phase. According to current research, synthetic data works best as a supplement rather than a replacement for human-generated data (Kang et al., 2025).

Assigning economic value to "free" data

LLMs and the applications they make possible "would not exist without having been trained on content taken from the digital commons: the ensemble of billions of open discussion forums, encyclopedia pages, digitized books and newspapers, repositories of open source code, and open access scientific publications, open digital images, videos, and sounds that are either in the public domain or available under a free and open license" (Noroozian et al., 2025). LLM builders typically seek to maximize the quality, size, and diversity of their training data, and the commons provides exactly what they need: vast, well-curated, openly licensed datasets that can be harvested without legal restriction (Mozilla Foundation, 2024). The characteristics of high-quality data in Table 1 align very closely with those of most curated collections.

Despite the obvious value of curated collections data for model training and inference, applying market economics to a freely available resource quickly gets messy. Curated collections function less like market goods and more like a common-pool resource: openly accessible, costly to produce and maintain, and relied upon by a range of stakeholders. While digital data can't be physically depleted, the capacity to steward it is rivalrous, bounded by funding, labour, and infrastructure. Large-scale AI harvesting increases

demands on the resource without directly or proportionally contributing back to its sustainability, and while AI companies derive significant commercial value from open data, the costs of curating and serving that data currently fall entirely on collection stewards. The concentration of the AI model training market within a few US technology companies that have enough resources to access and process large quantities of data, the training on public domain data (as well as data protected by copyright) underwrites a value transfer that further consolidates the power of the strongest players (Ivanova & Ding, 2025). Tarkowski and Warso (2024) describe this as a pattern of "free-riding" and "extraction of value from the commons" dating back at least a decade. Companies regard public web data as zero-cost input. The Wikimedia Foundation reports that, despite their recognition of the tremendous value of Wikimedia data, for tech companies, "The definition of value is cost to acquire the data — full stop" (Becker & Wyatt, 2023).

The question at the heart of current efforts is how to encourage proportional investment from AI companies in the open data providers that power their models without closing down the open internet or causing harm to the open curated collections community.

The sheer pace of AI development and the race to remain competitive have made thoughtful, proactive, or coordinated approaches to this challenge elusive. Open curated collections are looking for ways to either match the speed of the moment or force a slowdown — for example through regulation or litigation — that enables companies to stop and think.

Curated collections stewards could attempt to negotiate direct remuneration from the companies that harvest and use their content for AI training, but the incentives for companies to engage look very different when dealing with (relatively) small open datasets than when dealing with large rightsholding organizations like Newscorp. Open curated collections stewards typically have limited legal resources (and recourse) and asymmetric bargaining power. The market has yet to establish clear valuation models for non-commercial content in the context of AI harvesting. Consolidation in the media and publishing sector allows AI companies to strike deals that cover massive quantities of content, keeping transaction costs low (Sag, 2025). The newspapers included in OpenAI's deal with NewsCorp, for example, publish tens of thousands of new articles per month and have decades of archived material. The total corpus covered by the agreement is likely in the tens of millions of articles. While the largest sources of open data can command the

attention of model builders, smaller open collections stewards (representing the vast majority) face a serious problem of scale. Individual creators and small collections have little leverage in negotiations given the relative insignificance of any given data point to the effectiveness of an LLM (Chan et al., 2023).

Enterprise-grade API services, designed to support high-volume access, can provide a (potentially monetizable) gateway to open curated collections. The Model Context Protocol (MCP), released by Anthropic in 2024 and now considered an industry standard, allows content providers to set up a standardized endpoint that provides access to any MCP-compliant AI models, potentially making it easier for open curated collections to integrate with a wide range of AI tools (Tay, 2025). The importance of standardized and reliable access correlates with the increasing adoption of RAG, which relies on consultation of real-time data. Wikipedia, one of the two most-cited resources in ChatGPT outputs (Harsel, 2025), reported that its enterprise-grade API service successfully turned a profit in 2025 (Wikimedia Foundation, 2025), demonstrating that access plays a role alongside copyright or other legal concerns as an incentive for AI companies to compensate their most important data sources (Sag, 2025). It is unclear whether smaller open collections can capture enough attention and demonstrate sufficient return on investment to implement similar models.

Scraping fees collected via CloudFlare's pay-per-crawl model (Allen & Newton, 2025) or contracts granting collection-level scraping permissions governed by specific terms have also been floated as ways to remunerate smaller collections owners. Some new content licenses have been developed to explicitly support usage-based compensation. RSL, for example, allows users to express that a work is available via subscription, pay-per-crawl (compensation each time content is accessed), and pay-per-inference (compensation each time content contributes to an AI response).

However, even new licenses currently depend entirely on voluntary compliance. No major AI company has publicly committed to honouring RSL terms (Palmer, 2025), for example, and the protocol functions primarily as a request system that crawlers can ignore (Crowell & Moring, 2025). Enforcement is further complicated by the difficulty of tracking whether content appears in training datasets or influences model outputs (Brandom, 2025). So far, AI companies appear unmotivated to address the issues their practices create, and seem unwilling to adopt even minor mitigation measures such as rate-limiting their crawlers, let alone respecting creator preferences (Edwards, 2025). Evidence shows AI companies routinely ignore preference signals and consent frameworks absent legal consequences. Creative Commons received significant community pushback to its CC Signals framework,

specifically regarding the desire for more assertive opt-out terms and protections for creators (Ross, 2025).[10]

To succeed, these licensing frameworks need institutional or collective governance structures to pool resources, rights, and bargaining power from multiple content creators to negotiate with AI companies on their behalf. Negotiations, enforcement of terms, and payment processing could be managed by individual organizations or via community-based data trusts or intermediaries (Baack et al., 2025), which could include representatives from different stakeholder groups (i.e., industry, civil society organizations, curated collections stewards). Chan et al. (2023) propose the creation of a national public data trust to govern the use of the digital commons, negotiate for royalties on model revenues, and ensure compliance.

Initiatives like the RSL Collective (with 1,500+ member organizations as of December 2025) draw on collective licensing frameworks from music rights management (ASCAP, BMI) and academic publishing (Copyright Clearance Center). In the open collections sector, Sustainable Data Commons (SUDACO)[11] is developing a data trust mechanism based on "governance methodologies, and practical guides adapted to communities' needs, economic models, and social and ecological values." By aggregating individual rights through dedicated organizational structures, these approaches aim to reduce the burden on individual creators for monitoring, auditing, and enforcing rights while achieving sufficient scale to matter to AI companies who find negotiating with individual small rightsholders "administratively burdensome" (Baack et al., 2025). Through pooling resources, these organizations can potentially achieve sufficient scale and influence to matter to AI companies, reducing transaction costs that make individual negotiations prohibitive, and establishing reuse norms that change behaviour.

Collective structures face social and precedent barriers. AI companies currently access billions of tokens freely via platforms like Common Crawl, making voluntary payment for content they already obtain without restriction economically irrational without additional external pressures (Masnick, 2025; Mazzucato & Gernone, 2025). Building infrastructure to establish data governance, negotiate contracts, receive payments, and distribute royalties across member organizations requires substantial resources and time that many small collections lack. Additionally, collective rights organizations have histories of exploitative practices in other domains, raising governance concerns that require careful oversight (Band & Butler, 2013; Masnick, 2025).

[10] See, for example, the discussions in the CC Signals Github repository https://github.com/creativecommons/cc-signals/discussions.
[11] https://cis.cnrs.fr/en/sudaco-project/

Inviting AI companies into the commons

Beyond market-based licensing mechanisms, an alternative set of approaches focuses on strengthening the digital commons and establishing reciprocal norms for AI companies accessing openly shared content. The extant literature nearly uniformly advocates these approaches as an alternative or complement to market-based strategies for open curated collections. The commons model emphasizes collective coordination, social norm cultivation, and institutional trust-building rather than monetization of individual collections. Unlike licensing deals struck between AI firms and individual publishers, commons-based approaches seek to preserve the open, shared character of knowledge resources while ensuring that the entities profiting most from them contribute meaningfully in return. This perspective positions small collections not as individual vendors but as stewards of interconnected knowledge resources requiring collective protection (Verhulst et al., 2024).

The success of a commons-based approach assumes that AI companies do have incentives to engage more fairly with open collections stewards — incentives not predicated solely on legal risk. We describe several of these incentives below.

Surfacing shared incentives

- **Ensuring high-quality data sources remain online.** Excessive bot traffic and scraping threaten to push some small collections offline, as many lack the resources to "continue adding more servers, deploying more sophisticated firewalls, and hiring more operations engineers in perpetuity" (Weinberg, 2025; Grant, 2025). If key data sources go dark, AI companies lose access to the very content that makes their models useful.
- **Encouraging competition and discouraging consolidation.** Investment in the commons enables everyone, not only the wealthiest corporations, to build and refine models that lead to innovation. Even larger companies share a broad interest in encouraging innovation in the sector that they can benefit from in the future. In its announcement of its support for Harvard's Institutional Data Initiative, Microsoft cited the motivation to grow "a vibrant, competitive AI economy" by expanding access to the data resources needed to build LLMs (Davis, 2024). Openly licensed datasets can encourage competition and offer smaller players a way in. Adoption of the MCP (initially developed by Anthropic and later donated to the Linux Foundation) by the major AI market players is an example of industry-wide cooperation in this vein.

- **Retain scraping access.** The relationship between AI companies and the broader web is already showing signs of strain, and the consequences of ignoring this dynamic are already visible. A closing off of the web in response to AI crawlers, especially through blunt approaches that do not distinguish them from other machines, is affecting crawling for legitimate and widely accepted purposes, such as archiving and research. As of December 2025, around 5.6M websites had blocked OpenAI's GPTBot, a nearly 70% increase over the previous six months (Claburn, 2025). As scraping restrictions increase and more organizations adopt brute force approaches to thwarting bots, AI companies risk losing access to important sources of high-quality, novel training data.
- **Sustain high-quality, diverse training data.** Well-stewarded commons provide not just volume but also the metadata, documentation, and quality control that make training data more valuable. Companies may see value in building and sustaining resources that provide them with access to high-value, unique, or novel datasets. Some analyses have speculated that the open web will become polluted by low-quality, machine-generated content, making curated collections increasingly valuable data sources. Noroozian et al. (2025) write that AI model developers should have a vested interest in making curated collections data "identifiable, visible, and discoverable" in order to avoid 'model collapse' or increasingly more repetitive, biased, and less capable AI" caused by the growing presence of synthetic data across the web. Looking further ahead, Woahn (2026) predicts that "The next improvements in model capability will come from: highly specialized domain corpora; well-structured technical datasets; targeted refreshes rather than massive new ingestions; data with deep internal organization, not broad volume."
- **Fostering ongoing human contributions to the open web.** The commons is sustained by the continuous labour and ingenuity of human creators. Without new approaches for providing permission, credit, and compensation, these creators have diminishing incentives to openly share their work, and AI models lose access to original content (Chan et al., 2023, Huang & Siddarth, 2023). Borgman and Groth (2025) argue that scholars participate in a gifting economy in which they volunteer labour (such as sharing data) "with the expectation that these gifts create indebtedness, encourage reciprocity, and enhance reputations." To build trust among scholars and collections stewards, AI companies may need to more visibly and concretely adopt the norms of a gifting economy, for example, by ensuring proper attribution.
- **Creating a positive public image and consumer trust.** Beyond practical considerations, AI model developers may have a reputational incentive to demonstrate a commitment to "ethical" or "responsible" AI, which may include appropriate data harvesting practices. The non-profit Fairly Trained, for example, was launched for the purpose of certifying AI model developers and products that

adhere to certain standards for their training data (Knibbs, 2024). They also have an interest in serving consumers reliable information from robust sources in order to increase adoption and engagement with their platforms.

- **Mitigate regulatory and legal risks.** Finally, the legal landscape surrounding AI and data use remains in flux. Depending on the outcome of a number of lawsuits and pending legislation, AI model developers may need to fundamentally alter the way they harvest data. If they cannot rely on fair use justifications for scraping copyrighted data, for example, they will be increasingly reliant on openly licensed and public domain data. The strongest incentive for change could be future government policy that regulates the use of openly available data, for example, by strengthening creator opt-outs.

Despite these shared incentives, the challenge lies in bridging the gap between curated collections stewards and AI companies and developing the sociotechnical infrastructure needed to facilitate cross-sector engagement. Trust between commons communities and AI companies is severely eroded. Open source developer communities have expressed "deep frustration with what they view as AI companies' predatory behaviour toward open source infrastructure," undermining the relationship-building these approaches require (Edwards, 2025). Philosophically, there is tension between openness ideals and protection needs. The "open with thoughtfulness" paradigm (Metz, 2025) requires continuous judgment calls that may fragment the commons into incompatible governance zones.

Addressing these challenges will require deliberate effort on multiple fronts. The commons needs norms, governance frameworks, and contribution models developed with input from a range of stakeholders, including AI companies and technology platforms, as well as researchers, creators, and open curated collections stewards.

Building the infrastructure for reciprocity

A commons is sustained by well-designed contribution rules that are fair, observable, enforceable, and collectively governed. Working from the basis of shared incentives, open curated collections and AI companies can shift from a dynamic of asymmetric distribution of costs and benefits to one characterized by reciprocity: a bi-directional and mutually beneficial relationship where resources become "community assets" rather than targets for extracting knowledge (Tennison, n.d). Realizing this vision requires a fundamental shift in how we conceptualize data. Rather than treating information as "an object ripe for extraction," (Dulong de Rosnay & Stalder, 2020) we must begin to regard it as "a finite (and precious) resource that needs to be nurtured and cared for" (Verhulst et al., 2024).

Establishing norms, the shared expectations that guide participation and reciprocity in the commons, is the first layer of coordination. Norms may include respecting creator preferences and contributing back to the ecosystem, financially or otherwise, proportionally to the pressure you place on it. They may also cover specific technical and operational guidelines such as provenance tracking, required attribution, restrictions on verbatim reproduction, or barriers against training new foundational models. They may be supported by technical infrastructure. For example, Padilla (2026) contends that the MCP can help to facilitate provenance tracking and attribution. Norms can function as an effective and more values-aligned alternative to stronger intellectual property rights or individually negotiated contracts. Their power "comes from coordination and solidarity ... across sectors, communities, and geographies" (Hardinges, et al., 2025).

Norms spread faster and carry more weight when embedded in structures capable of enforcing and evolving them. Commons governance frameworks establish coordinated rules and institutional structures for managing shared data resources, addressing what Tarkowski and Warso (2023) identify as a "governance vacuum in the process of creating and using datasets for AI training."

Norms and governance provide the principles; contribution frameworks provide the mechanism.

Commons-based contribution models do not aim to price the data itself, but to ensure that contributions align with the real costs of sustaining the commons. Those costs are substantial and varied, encompassing labour (including curation, metadata creation, quality control, governance, and community engagement), infrastructure (hosting, storage, and bandwidth), and long-term stewardship responsibilities such as preservation and migration to new formats and platforms.

Rather than charging per record or per byte of data harvested — a transactional model that can feel extractive and that poorly captures the true dynamics of commons use — contributions are structured to reflect how much pressure an actor places on the system. This includes the volume of access (such as scraping frequency, API usage, or downloads) and the exclusivity of value capture (for example, different standards might apply to open versus proprietary models). The result is a contribution model expressed as a weighted share of total costs, not a one-time fee. Contributions may include monetary payments as well as non-monetary value return mechanisms. For commons-oriented collections philosophically opposed to commodifying knowledge, reciprocal contributions — improved metadata, processing tools, derived datasets — may align better with mission than licensing revenue. The Common Pile team, for instance, has demonstrated this principle in practice

by sharing improved transcripts and metadata with the source platforms from which their training data was drawn (Baack et al., 2025).

Contribution models must take into account the real costs of keeping curated collections viable, including labour (curation, metadata creation, quality control, governance, community engagement), infrastructure (hosting, storage, bandwidth), and long-term stewardship (digital preservation, migrations).

Under this framework, commons stakeholders bear responsibilities proportional to both their impact on the commons and their capacity to contribute. Rather than flowing through one-off bilateral payments, contributions are directed into collective sustainability funds and community trusts. This structure avoids the power imbalances inherent in direct negotiation and supports the kind of collective governance that the commons requires. Non-monetary contributions, such as code, tooling, or documentation, are valued, but do not replace the need for financial support.

One of the most direct ways AI companies can contribute to the commons is by funding the human labour that keeps it alive. Mike Masnick (2024) has proposed a version of this idea in the context of journalism: "If the tech companies need good, well-written content to fill their training systems, and the world needs good, high-quality journalism, why don't the big AI companies agree to start funding journalists and solve both problems in one move?" Mazzucato & Gernone (2025) similarly argued for an increase in public funding for artists and creators paid for via taxes levied on AI companies.

In the context of open curated collections, this might look like funding the digitization and curation of unique, high-quality, well-structured content optimized for AI use. Investing in thoughtfully curated collections reduces reliance on the unfiltered and unstructured datasets like Common Crawl, which currently dominate LLM training, resulting in improved transparency and performance (Mozilla Foundation, 2025). AI companies could invest in the creation of Compute-Ready Documents (CRD) "designed explicitly for agentic ingestion rather than blind chunking," ensuring that training data is not only abundant but structurally suited to the systems that will use it (Gunter, 2026).

Sustaining the commons also requires investment in shared institutional infrastructure.

Verhulst (2025) has proposed the creation of trusted institutions — modelled on libraries — that would steward equitable and responsible access to data for the AI era. Current examples of this approach include CodeCommons from Software Heritage and Harvard Law School Library's Institutional Data Initiative.[12] These examples demonstrate ways in which trusted institutions such as libraries and archives can build high-quality corpora, establish norms and guidelines around reuse, and bridge different stakeholders in the AI data economy.

Conclusion

Curated collections face sustainability challenges driven by automated data harvesting. These repositories provide high-quality, well-structured, authoritative data that AI companies need for training foundation models and supporting generation systems. However, bot traffic creates costs through bandwidth consumption and disruptions to service, while AI intermediaries divert human engagement. Current mitigation strategies present significant limitations: blocking crawlers risks fragmenting the open web, preference signals depend on voluntary compliance that is frequently ignored by crawlers, and legal frameworks vary across jurisdictions and provide patchwork systems based on outdated policies. Defensive responses that restrict access may accelerate consolidation of data resources only to well-resourced corporations.

Stark scale disparities are evident in the emergent initiatives like the licensing market. Major content providers secure substantial multi-year deals, with news organizations and large publishers engaging in agreements that net them millions of dollars. These agreements reflect AI companies' need for reliable access, clear rights, and documented provenance. Small collections, on the other hand, face structural barriers; AI companies seem to prefer aggregated deals that reduce transaction costs, which means individual small collections lack negotiating leverage. Collective licensing initiatives like the RSL Collective attempt to address these challenges by aggregating collections to achieve scale, drawing on models from music rights management and academic publishing. Machine-readable protocols provide technical infrastructure for expressing usage terms, while frameworks like CC

[12] https://codecommons.org/, https://hls.harvard.edu/library/about-the-library/library-staff/institutional-data-initiative/

Signals enable contextual preference specification. However, major AI companies have not committed to honouring such approaches and enforcement mechanisms remain limited.

Alternative approaches emphasize commons governance and reciprocal norms rather than individual monetization. Commons governance frameworks could establish coordinated rules for managing shared resources and addressing power asymmetries, while recognizing potential incentives for AI companies, such as maintaining access to high-quality data sources and mitigating regulatory risks. Contribution models could align payments with actual stewardship costs like curation labour, infrastructure resources, and long-term preservation.

The fundamental challenge remains in preserving commons principles while adapting to unprecedented technological demands. This landscape scan surveys emergent mechanisms to sustain openly accessible resources while addressing extractive practices in a meaningful way.

Acknowledgements

We thank our reviewers (Bridget Almas, Arran Griffith, Ann Michael, and Thomas Padilla) for their insights and comments on the draft of this manuscript.

This publication is supported by the Mellon Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect opinions or policies of the Mellon Foundation.

AI use statement

The authors of this report used AI-driven tools, including ChatGPT, Google Scholar Labs, and Claude, to identify some sources cited in this report. All sources were verified and reviewed by the authors. The authors drafted all content for this report and used AI-driven tools in some sections to revise language for improved clarity and readability.

Contributors

Bridget Almas, https://orcid.org/0000-0001-7556-1572: writing – review & editing
Lauren Collister, https://orcid.org/0000-0001-5767-8486: conceptualization, methodology, investigation, formal analysis, writing – original draft, writing – review & editing
Emma Green, https://orcid.org/0000-0003-3620-6855: conceptualization, methodology, investigation, supervision, writing – review & editing
Arran Griffith, https://orcid.org/0009-0001-8530-6214: writing – review & editing
Sarah Lippincott, https://orcid.org/0000-0002-5700-5844: conceptualization, methodology, investigation, formal analysis, writing – original draft, writing – review & editing
Ann Michael, https://orcid.org/0000-0003-3332-1970: writing – review & editing
Thomas Padilla, https://orcid.org/0000-0002-6743-6592: writing – review & editing
Katherine Skinner, https://orcid.org/0000-0003-0139-7524: conceptualization, methodology, investigation, supervision, writing – review & editing
Kaitlin Thaney, https://orcid.org/0000-0002-7217-4494: conceptualization, investigation, writing – review & editing
Emmy Tsang, https://orcid.org/0000-0002-9248-1280: writing – review & editing
Chrys Wu, https://orcid.org/0000-0002-8431-1580: investigation, writing – review & editing

Appendix A: Inventory of selected licensing deals

- Multiyear content licensing deal between OpenAI and News Corp “worth over $250M” (reported as a five-year arrangement). Includes cash payment as well as credits to use OpenAI technology (Bruell et al., 2024).
- OpenAI is paying Dotdash Meredith at least $16M a year to license its content for AI (David, 2024a).
- Microsoft paid Informa (parent company of Taylor & Francis) an initial access fee “around $10M” with recurring payments across the following years for scholarly/journal content (Potter, 2024). Taylor & Francis’ broader 2023 licensing income (reportedly around $75M) suggests multiple deals possibly covering multiple AI clients (Lewin, 2025).
- John Wiley & Sons reported an anticipated $44M in revenue from AI licensing deals with at least two undisclosed partners (Lewin, 2025).
- Shutterstock disclosed its AI-licensing business generated approximately $104 million in 2023 and forecast higher numbers; it has multi-year arrangements (including a six-year agreement with OpenAI) (Ford, 2024).
- In some instances, deals are structured on a per-item basis rather than for a full corpus. For example, Paul & Tong (2024) report that PhotoBucket has “discussed rates of between 5 cents and $1 per photo and more than $1 per video” for access to their archive of 13 billion photos and videos. The authors also cite Freepik’s licensing of the majority of its archive of 200 million images at 2 to 4 cents per image to two major AI companies. A data aggregator that licenses data to a range of major companies reported the market rate for images at $1 to $2, $2 to $4 for short-form video and $100 to $300 per hour for longer films, and $0.001 per word for text.
- Reddit receives an estimated $60M annually from Google to use its data for AI purposes (Brandom, 2025).
- Content providers are also introducing enterprise API access models rather than licensing deals. Wikimedia Enterprise, for example, reported that its enterprise API service netted a profit of $646k between 2022 and 2025. The service’s total revenue for the 2024-2025 fiscal year was $8.3M (Wikimedia Foundation, 2025).
- OpenAI donated $50 million to 15 research institutions, including Harvar,d as part of its NextGenAI consortium project (Chen & Im, 2025). That includes supporting the

nearly 1 million public-domain books Harvard plans to release for AI training as well as many other large-scale initiatives around AI at the partner research institutions.

- Anthropic gave $1.5M to the Python Software Foundation to improve security, in exchange for access to code (Crary, 2026).
- Anthropic recently announced integrations with bioRxiv, medRxiv, and ClinicalTrials.gov; the press release did not include details about any potential compensation (Advancing Claude in Healthcare and the Life Sciences, 2026).

Content Provider	AI Company	Reported Financial Value	Length/Duration
Associated Press	OpenAI	Undisclosed	2023
The Atlantic	OpenAI	Undisclosed	Multi-year
Axel Springer	OpenAI	Undisclosed	Multi-year
Dotdash Meredith	OpenAI	Undisclosed	Multi-year
Financial Times	OpenAI	Undisclosed	Multi-year
Le Monde	OpenAI	Undisclosed	Multi-year
News Corp	OpenAI	$250M total	5 years
Prisa Media (El País)	OpenAI	Undisclosed	Multi-year
Reddit	OpenAI	Undisclosed	Multi-year
Shutterstock	OpenAI	$25–50M	6 years
Stack Overflow	OpenAI	Undisclosed	Multi-year
Thomson Reuters	OpenAI	Undisclosed	Multi-year
Quora (Poe data)	OpenAI	Undisclosed	2024
Reddit	Google	$60M per year	Long-term
Stack Overflow	Google (Gemini)	Undisclosed	2024
Shutterstock	Google	Undisclosed	2023 extension
The Atlantic	Meta	Undisclosed	2024
Shutterstock	Meta	Undisclosed	2022 extension
Informa (Taylor & Francis)	Microsoft	$10M (confirmed)	2023
Westlaw/Thomson Reuters	Microsoft	Undisclosed	Ongoing
Pearson Education	Microsoft (rumoured)	Undisclosed	Undisclosed

LexisNexis	Microsoft, OpenAI	Undisclosed	Ongoing
Getty Images	NVIDIA	Undisclosed	2023
Yelp	Perplexity	Revenue-share (no fixed amount)	Ongoing
Cengage	Various	Undisclosed	Ongoing
Elsevier	Various	Undisclosed	Ongoing
Informa (Taylor & Francis)	Various	$75M aggregate revenue	2023
Photobucket	Various	$4–5M total	2023
Springer Nature	Various	Undisclosed	Ongoing
X (Twitter)	Various	$0.5M–$2.5M per year	Annual

Table 2. Publicly disclosed deals between content providers and AI companies. Table generated by ChatGPT based on data collected by Émet Research.

Works Cited

Advancing Claude in healthcare and the life sciences. (2026, January 11). https://www.anthropic.com/news/healthcare-life-sciences

Allen, W., & Newton, S. (2025, July 1). *Introducing pay per crawl: Enabling content owners to charge AI crawlers for access*. The Cloudflare Blog. https://blog.cloudflare.com/introducing-pay-per-crawl/

Baack, S., Biderman, S., Odrozek, K., Skowron, A., Bdeir, A., Bommarito, J., Ding, J., Gahntz, M., Keller, P., Langlais, P.-C., Lindahl, G., Majstorovic, S., Marda, N., Penedo, G., Segbroeck, M. V., Wang, J., Werra, L. von, Baker, M., Belião, J., … Wolf, T. (2025). *Towards Best Practices for Open Datasets for LLM Training* (No. arXiv:2501.08365). arXiv. https://doi.org/10.48550/arXiv.2501.08365

Biderman, S., Majstorovic, S., & Skowron, A. (2025, June 5). *The Common Pile v0.1*. EleutherAI Blog. https://blog.eleuther.ai/common-pile/

bioRxiv. (2025, November 4). *Q.e.d Science – an AI review tool for authors*. https://connect.biorxiv.org/news/2025/11/04/qed_review_tool

Becker, L., & Wyatt, L. (2023, August 19). *What We Have Learned from Large Commercial Users*. Wikimania Singapore, Singapore. https://upload.wikimedia.org/wikipedia/commons/2/2b/What_we_have_learned_from_talking_to_large_commercial_users_of_Wikimedia_content.pdf

Boboris, K. (2025, October 31). Attention Authors: Updated Practice for Review Articles and Position Papers in arXiv CS Category – arXiv blog. *arXiv Blog*. https://blog.arxiv.org/2025/10/31/attention-authors-updated-practice-for-review-articles-and-position-papers-in-arxiv-cs-category/

Borgman, C. L., & Groth, P. (2025). Infrastructure, Intermediaries, and Artificial Intelligence: A Rejoinder to Commentaries on “From Data Creator to Data Reuser: Distance Matters". *Harvard Data Science Review*, *7*(2). https://doi.org/10.1162/99608f92.c17c3adb

Brandom, R. (2025, September 10). RSS co-creator launches new protocol for AI data licensing. *TechCrunch*. https://techcrunch.com/2025/09/10/rss-co-creator-launches-new-protocol-for-ai-data-licensing/

Browne, R. (2024, August 1). *World’s first major AI law enters into force — Here’s what it means for U.S. tech giants*. CNBC. https://www.cnbc.com/2024/08/01/eu-ai-act-goes-into-effect-heres-what-it-means-for-us-tech-firms.html

Bruell, A., Schechner, S., & Seetharaman, D. (2024, May 22). OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over $250 Million. *Wall Street Journal*. https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba

Chan, A., Bradley, H., & Rajkumar, N. (2023, March 16). *Reclaiming the Digital Commons: A Public Data Trust for Training Data*. arXiv.Org. https://arxiv.org/abs/2303.09001v2

Chen, S., & Im, D. J. (2025, March 10). OpenAI Donates $50 Million for AI Use in Research at Harvard, 14 Other Institutions. *The Harvard Crimson*. https://www.thecrimson.com/article/2025/3/11/openai-50mil-funding/

Claburn, T. (2025, December 8). Publishers say no to AI scrapers, block bots at server level. *The Register*. https://www.theregister.com/2025/12/08/publishers_say_no_ai_scrapers/

Courtney, K. (2024, March 28). All TDM & AI Rights Reserved? Fair Use & Evolving Publisher Copyright Statements. *SPARC*. https://sparcopen.org/news/2024/all-tdm-ai-rights-reserved/

Crary, L. (2026, January 13). Python Software Foundation News: Anthropic invests $1.5 million in the Python Software Foundation and open source security. *Python Software Foundation News*. https://pyfound.blogspot.com/2025/12/anthropic-invests-in-python.html

David, E. (2024, May 7). *OpenAI strikes licensing deal with the magazine giant behind People*. The Verge. https://www.theverge.com/2024/5/7/24151171/openai-dotdash-meredith-people-instyle-licensing

Davis, B. (2024, December 12). Supporting New Open Data Initiatives: Harvard's Institutional Data Initiative and CORE. *Microsoft On the Issues*. https://blogs.microsoft.com/on-the-issues/2024/12/12/supporting-new-open-data-initiatives-institutional-data-initiative-and-core/

Davis, C. (2025). Retrieval-Augmented Generation for Web Archives: A Comparative Study of WARC-GPT and a Custom Pipeline. *The Code4Lib Journal*, (61). https://journal.code4lib.org/articles/18555

Delaney, J. (2025, November 3). Answer Engines Redefine Search. *Communications of the ACM*. https://cacm.acm.org/news/answer-engines-redefine-search/

del Rio-Chanona, R. M., Laurentsyeva, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, *3*(9), pgae400. https://doi.org/10.1093/pnasnexus/pgae400

Dulong de Rosnay, M., & Stalder, F. (2020). Digital commons. *Internet Policy Review, Concepts of the Digital Society*, *9*(4). https://doi.org/10.14763/2020.4.1530

Edwards, B. (2025, March 25). Open source devs say AI crawlers dominate traffic, forcing blocks on entire countries. *Ars Technica*. https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/

Eve, M. P. (2025, November 24). Creative commons licenses and copyright may not stop academic work being used to train AI. *Impact of Social Sciences - Maximizing the Impact of Academic Research*. https://blogs.lse.ac.uk/impactofsocialsciences/2025/11/24/creative-commons-licenses-and-copyright-may-not-stop-academic-work-being-used-to-train-ai/

Ford, B. (2024, June 4). Shutterstock's AI-Licensing Business Generated $104 Million Last Year. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2024-06-04/shutterstock-s-ai-licensing-business-generated-104-million-last-year

Grant, S. (2025, June 5). *Keeping the Web Up Under the Weight of AI Crawlers*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2025/06/keeping-web-under-weight-ai-crawlers

Greaves, S. (2025, December 9). AI calls time on curation [Substack newsletter]. *Scholarly Futures*. https://scholarlyfutures.substack.com/p/ai-calls-time-on-curation

Gunter, D. (2026, February 2). Scholarly publishing's great leap. *Research Information*. https://www.researchinformation.info/analysis-opinion/scholarly-publishings-great-leap/

Hardinges, J. (2025, December 1). No, AI Hasn't Run Out of Data... Here's What's Really Happening. *OpenMined*. https://openmined.org/blog/ai-hasnt-run-out-of-data/

Hardinges, J., Pearson, S., & Ross, R. (2025). *From Human Content to Machine Data: Introducing CC Signals* (p. 34). Creative Commons. https://creativecommons.org/wp-content/uploads/2025/06/Human-Content-to-Machine-Data_Final.pdf

Hellman, E. (2025, March 21). AI bots are destroying Open Access. *Go to Hellman*. https://go-to-hellman.blogspot.com/2025/03/ai-bots-are-destroying-open-access.html

Holscher, E. (2024, July 25). *AI crawlers need to be more respectful*. Read the Docs. https://about.readthedocs.com/blog/2024/07/ai-crawlers-abuse/

Huang, S., & Siddarth, D. (2023). *Generative AI and the Digital Commons* (arXiv:2303.11074). arXiv. https://doi.org/10.48550/arXiv.2303.11074

Intelligence artificielle et collections de la BnF: L'exemple de l'HTR (Handwritten Text Recognition). (n.d.). BnF - Site institutionnel. Retrieved January 20, 2026, from https://www.bnf.fr/fr/intelligence-artificielle-et-collections-de-la-bnf-lexemple-de-lhtr-handwritten-text-recognition

Ivanova, V., & Ding, J. (2025). *Choral Data "Trust" Experiment White Paper*. Serpentine Arts Technologies. https://doi.org/10.5281/zenodo.14859320

Kang, F., Ardalani, N., Kuchnik, M., Emad, Y., Elhoushi, M., Sengupta, S., Li, S.-W., Raghavendra, R., Jia, R., & Wu, C.-J. (2025). *Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls*. https://arxiv.org/html/2510.01631v1

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Knibbs, K. (2024, January 17). This Tech Exec Quit His Job to Fight Generative AI's Original Sin. *Wired*. https://www.wired.com/story/ai-executive-ed-newton-rex-turns-crusader-stand-up-for-artists/

Lewin, F. (2025, June 26). *Market Report: AI Data Licensing Deals (2020 - Present)*. Émet Research. https://www.emetresearch.ai/blogs/market-report-ai-data-licensing-deals-(2020-present)

Lin, J., Shan, R., Zhu, J., Xi, Y., Yu, Y., & Zhang, W. (2025). *Stop DDoS Attacking the Research Community with AI-Generated Survey Papers* (No. arXiv:2510.09686). arXiv. https://doi.org/10.48550/arXiv.2510.09686

Lloyd, T. (2026). "Access in Flux: Three Hot Topics You Need to Know About." NISO Plus 2026 (Feb 16-18, 2026).

Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., ... Pentland, S. (2024). *Consent in Crisis: The Rapid Decline of the AI Data Commons* (No. arXiv:2407.14933). arXiv. https://doi.org/10.48550/arXiv.2407.14933

Loring, J., & Rayner, W. (2025, October 29). *Anthropic Settlement Resets Balance of Power for Content Creators*. Bloomberg Law. https://news.bloomberglaw.com/legal-exchange-insights-and-commentary/anthropic-settlement-resets-balance-of-power-for-content-creators

Malhotra, S. (2025, September 16). The Anthropic Settlement That Wasn't: Copyright Battles. *Ronin Legal*. https://roninlegalconsulting.com/the-anthropic-settlement-that-wasnt/

Masnick, M. (2024, April 11). "An Only Slightly Modest Proposal: If AI Companies Want More Content, They Should Fund Reporters, And Lots Of Them." *Techdirt*. https://www.techdirt.com/2024/04/11/an-only-slightly-modest-proposal-if-ai-companies-want-more-content-they-should-fund-reporters-and-lots-of-them/

Masnick, M. (2025, September 8). We're Walling Off The Open Internet To Stop AI — And It May End Up Breaking Everything Else. *Techdirt*. https://www.techdirt.com/2025/09/08/were-walling-off-the-open-internet-to-stop-ai-and-it-may-end-up-breaking-everything-else/

Matas, L. (2025). *Enhancing Visibility Across Languages: Semantic Multilingual Search for Scholarly Content* (Version 1). Confederation of Open Access Repositories (COAR). https://doi.org/10.5281/zenodo.17535337

Mazzucato, M., & Gernone, F. (2025, July 25). *AI Should Help Fund Creative Labor*. Project Syndicate. https://www.project-syndicate.org/onpoint/how-ai-profits-can-help-fund-cultural-production-by-mariana-mazzucato-and-fausto-gernone-2025-07

Metz, R. (2026, January 9). Openness Has Limits [Substack newsletter]. *The Digital Shift*. https://rosalynmetz.substack.com/p/openness-has-limits?utm_medium=android&triedRedirect=true

Miller, M. *AI Scrapers vs Wikibase*. https://semlab.io/blog/ai-scrapers-vs-wikibase.html

Monthly Submissions. (n.d.). arXiv.Org. Retrieved March 10, 2026, from https://arxiv.org/stats/monthly_submissions

Mozilla Foundation. (2024, February). Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

Mueller, B., Foundation, W., Danis, C., Lavagetto, W. F. G., & Foundation, W. (2025, April 1). How crawlers impact the operations of the Wikimedia projects. *Diff*. https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/

Mulvany, I. (2025, March 25). *AI Bot traffic — A real problem, right now.* https://world.hey.com/ian.mulvany/ai-bot-traffic-a-real-problem-right-now-a6a513a3

Noroozian, A., Aldana, L., Arisi, M., Asghari, H., Avila, R., Bizzaro, P. G., Chandrasekhar, R., Consonni, C., Angelis, D. D., Chiara, F. D., Rio-Chanona, M. del, Rosnay, M. D. de, Eriksson, M., Font, F., Gomez, E., Guillier, V., Gutermuth, L., Hartmann, D., Kaffee, L.-A.,

... Wasielewski, A. (2025). *Generative AI and the Future of the Digital Commons: Five Open Questions and Knowledge Gaps* (No. arXiv:2508.06470). arXiv. https://doi.org/10.48550/arXiv.2508.06470

Padilla, T. (2026, February 16). Maintaining Truth in a Generative AI Era [Substack newsletter]. *Memory Work*. https://thomaspadilla.substack.com/p/maintaining-truth-in-a-generative

Pasetti, M., Santos, J. W., Corrêa, N. K., De Oliveira, N., & Barbosa, C. P. (2025). Technical, legal, and ethical challenges of generative artificial intelligence: An analysis of the governance of training data and copyrights. *Discover Artificial Intelligence*, *5*(1), 193. https://doi.org/10.1007/s44163-025-00379-6

Paul, K., & Tong, A. (2024, April 5). Inside Big Tech's underground race to buy AI training data. *Reuters*. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/

Penti, R. S., & Schaal, Y. (2025, September 8). Anthropic's Landmark Copyright Settlement: Implications for AI Developers and Enterprise Users. *Ropes & Gray LLP*. https://www.ropesgray.com/en/insights/alerts/2025/09/anthropics-landmark-copyright-settlement-implications-for-ai-developers-and-enterprise-users

Potter, W. (2024, July 23). *An academic publisher has struck an AI data deal with Microsoft — without their authors' knowledge*. The Conversation. https://doi.org/10.64628/AA.aggyxe33a

Prater, S., Wrobel, T., Gillmore, J., & Metz, R. (2025). *Aggressive AI Harvesting of Digital Resources*. LYRASIS Wiki. https://wiki.lyrasis.org/pages/viewpage.action?pageId=364743621

Repositories in the Age of AI: The Attack of the Bots. (2025, October 30). https://www.youtube.com/watch?v=KY0dhvbJYNA

Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper No. 4438593). Social Science Research Network. https://doi.org/10.2139/ssrn.4438593

Sag, M. (2025, November 19). The False Hope of Content Licensing at Internet Scale. *ProMarket*. http://www.promarket.org/2025/11/19/the-false-hope-of-content-licensing-at-internet-scale/

Schaffer, A., Oremus, W., & Tiku, N. (2026, January 27). Inside one company's secret plan to 'destructively scan every book in the world.' *The Washington Post*. https://www.washingtonpost.com/technology/2026/01/27/anthropic-ai-scan-destroy-books/

Shearer, K. (2025). *COAR Strategic Analysis of the Scholarly Communications Environment*. Confederation of Open Access Repositories. https://coar-repositories.org/wp-content/uploads/2025/11/COARs-Strategic-Analysis-of-the-Landscape-2025-3.pdf

Shearer, K., & Walk, P. (2025). *The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025*. Confederation of Open Access Repositories. https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

Tarkowski, A., & Warso, Z. (2023). *AI_COMMONS: Filling the governance vacuum related to the use of information commons for AI training*. Open Future. https://openfuture.eu/wp-content/uploads/2023/01/ai-commons-report.pdf

Tarkowski, A., & Warso, Z. (2024). Commons-based Data Set Governance for AI. *Open Future*. https://openfuture.pubpub.org/pub/principles-for-commons-based-data-set-governance-for-ai/release/4

Tay, A. (2025, November 30). Model Context Servers — Wiley AI Gateway & PubMed — How Claude can now pilot test search strategies using PubMed [Substack newsletter]. *Aaron Tay's Musings about Librarianship*. https://aarontay.substack.com/p/mcp-servers-and-academic-search-the

Tennison, J. (n.d.). *Creative communities*. Open Future. Retrieved January 22, 2026, from https://openfuture.eu/paradox-of-open-responses/creative-communities

Terras, M. (2025). Be More Library: Upholding Library Values in a Tech Industry World. In S. Oehlschläger & A. Wenzel (Eds.), *Artificial Intelligence Meets Cultural Heritage: The Transformative Power of AI for and through National Libraries*. Conference of European National Librarians (CENL). https://www.cenl.org/wp-content/uploads/2026/01/ArtificialIntelligenceMeetsCulturalHeritage.pdf

Van Noorden, R. (2020). The ethical questions that haunt facial-recognition research. *Nature*, *587*(7834), 354–358. https://doi.org/10.1038/d41586-020-03187-3

Verhulst, S. (2025, August 19). *Why We Need a Carnegie Moment for the Age of AI* Tech Policy Press. https://techpolicy.press/why-we-need-a-carnegie-moment-for-the-age-of-ai

Verhulst, S., Chafetz, H., & Zahuranec, A. J. (2024). *Data Commons: Under Threat by or The Solution for a Generative AI Era? Rethinking data access and re-use* (SSRN Scholarly Paper No. 4836354). Social Science Research Network. https://doi.org/10.2139/ssrn.4836354

Vigliarolo, B. (2026, February 4). AI bot traffic closing in on human web visits, study finds. *The Register*. https://www.theregister.com/2026/02/04/ai_bot_traffic_web_browsers/

Wang, J. T., & Jia, R. (2023). Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 6388–6421. https://proceedings.mlr.press/v206/wang23e.html

Weinberg, Michael. (2025, June). *Are AI Bots Knocking Cultural Heritage Offline?* GLAM-E Lab. https://glamelab.org/products/are-ai-bots-knocking-cultural-heritage-offline/

Weiß, E.-M. (2026, January 28). OpenStreetMap is concerned: Thousands of AI bots are collecting data | heise online. *Heise Online*. https://www.heise.de/en/news/OpenStreetMap-is-concerned-thousands-of-AI-bots-are-collecting-data-11157359.html

Wiggers, K. (2024, June 1). AI training data has a price tag that only Big Tech can afford. *TechCrunch*. https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/

Wiggers, K. (2024, March 13). OpenAI's deals with publishers could spell trouble for rivals. *TechCrunch*. https://techcrunch.com/2024/03/13/are-openais-deals-with-publishers-edging-out-the-competition/

Wikimedia Foundation. (2025, November 24). *Wikimedia Enterprise Financial Report: Fiscal Year 2024 – 2025*. https://diff.wikimedia.org/2025/11/24/wikimedia-enterprise-financial-report-fiscal-year-2024-2025/

Woahn, J. (2026, January 21). Guest Post: AI Isn't Going to Pay for Content ... At Least Not How You're Hoping It Will. *The Scholarly Kitchen*. https://scholarlykitchen.sspnet.org/2026/01/21/guest-post-ai-isnt-going-to-pay-for-content-at-least-not-how-youre-hoping-it-will/

Wu, Y. (2023, July 27). China's Interim Measures to Regulate Generative AI Services: Key Points. *China Briefing News*. https://www.china-briefing.com/news/how-to-interpret-chinas-first-effort-to-regulate-generative-ai-measures/

Xu, Y. (2026, January 8). Anticircumvention Law is Not the Right Solution to Webscraping. *Authors Alliance*. https://www.authorsalliance.org/2026/01/08/anticircumvention-law-is-not-the-right-solution-to-webscraping/