

Are Bots Re-Shaping Open Access?

by Britt Amell | 7 October 2025 | English, Insights and Signals Reports



Lisez-le en français

This insights and signals report was written by Brittany Amell, with thanks to INKE Partner James MacGregor (Canadian Research Knowledge Network) for the comments and contributions.

At a Glance / En un coup d'œil

Topic / Titre	AI bots, open scholarship, open infrastructure
Key Participants / Créateur	Coalition of Open Access Repositories, Canadian Research Knowledge Network / Réseau canadien de documentation pour la recherche, Internet Archive

Date / Période	2025
Keywords / Mots-clés	AI safety / Sécurité de l'IA, AI bots / Robots d'indexation IA, AI governance / Gouvernance de l'IA, open access / libre accès, open infrastructure / infrastructure ouverte, open social scholarship / approches sociales des savoirs ouverts, generative artificial intelligence / l'intelligence artificielle générative

Summary

Modern AI training bots are overwhelming open access repositories worldwide, potentially forcing institutions to choose between protecting their infrastructure and maintaining open principles. This insights and signals report offers a brief introduction to ‘bots,’ some of the issues they pose, as well as some early responses from the open access and open scholarship community.

The Evolution of Web Bot Behavior

The landscape of automated web traffic has fundamentally changed over the past 25 years. Traditionally, it seems web crawlers operated with a degree of digital etiquette—they respected robot exclusion protocols (or robots.txt) that communicated which parts of a site they should or should not access, identified themselves clearly, and maintained reasonable request rates (Hellman 2025; Weinberg 2025). Sometimes, a bot might spam a site with registrations, but administrators could block them fairly easily based on IP address—they were, as Hellman (2025) puts it, “part of the landscape, not the dominant feature.”

These days, however, they don’t seem to make bots like they used to.

Modern AI bots demonstrate what Hellman (2025) describes as “mindless” behavior, utilizing maximum available server connections and ramping up requests when additional capacity becomes available. Unlike their predecessors, these bots often use randomized user-agent strings, operate from large IP address blocks, and

can become trapped in endless loops, making thousands of requests for non-functional links. In many respects, misbehaving bot behaviour is indistinguishable from those of Distributed Denial of Services (DDOS) attacks, where the attacker intentionally uses spoofing practices and automated access requests to overwhelm a target web server. This aggressive approach has led to significant resource consumption, causing some systems to slow down and others to crash completely.

Technical and Operational Challenges

The impact on open access repositories has been substantial. The [2025 survey](#) from the Coalition of Open Access Repositories (COAR) revealed that over 90% of the 66 responding repositories worldwide experienced problems with aggressive bots, often multiple times per week, resulting in performance degradation or complete service outages ([Shearer and Walk 2025](#)). These disruptions force repository staff who are already stretched thin to scramble as they try to find a solution that allows them to implement protective measures while maintaining open access principles.

Repositories exist to make knowledge available and useful, but when aggressive bots continue to cause problems, they may be forced to limit access to their resources.

The phenomenon of “swarming” has emerged as a particular concern. As [Weinberg \(2025\)](#) explains, this involves large numbers of bots visiting a collection simultaneously, downloading every available thing and following every discoverable link. This behaviour differs from typical human users, who tend to focus on specific content areas.

By attempting to harvest everything indiscriminately and at once, these bots create spikes of traffic that are unpredictable and overwhelming for a repository’s server.

Defensive Measures and Their Limitations

Repository staff have attempted to mitigate and fend off the bot attacks, with varying

degrees of success ([Panitch 2025](#)). However, while some of these measures may be successful at blocking bots, “it is also clear they are impeding access to the repositories by other more welcome actors, such as individual humans and benign systems,” write Shearer and Walk (2025, p. 1) in their [report for COAR](#).

For instance, many have turned to commercial services like Cloudflare to block bots, but as [Hellman \(2025\) notes](#), this has had repercussions for “good” bots: “Internet Archive can no longer save snapshots of one of the best open-access publishers, MIT Press because of Cloudflare blocking.”

The effectiveness of traditional protocols like robots.txt has also diminished. While this voluntary compliance system worked well with earlier generations of web crawlers, many AI training bots now ignore these requests entirely—representing a departure from established internet norms and etiquette: “The [robots.txt] protocol has not proven to be as effective in the context of bots building AI training datasets. Respondents reported that robots.txt is being ignored by many (although not necessarily all) AI scraping bots. This was widely viewed as breaking the norms of the internet, and not playing fair online” (Weinberg 2025).

Legal and Licensing Considerations

The current licensing framework for open access content is also facing new challenges in the AI era, has brought on an insatiable need for more training data, writes Decker in a [recent guest post for The Scholarly Kitchen](#). Most open access academic content uses Creative Commons licenses, particularly CC-BY, but as Decker (2025) points out, these licenses were designed with human readers and traditional reuse scenarios in mind.

The massive scale and appetite for AI training data doesn’t neatly fit into traditional categories of copying, distribution, or adaptation, says [Decker \(2025\)](#):

When AI models ingest text, this is disaggregated into ‘tokens’ (i.e., words) that are transformed into neural networks, which in turn form the basis for query responses. Text is converted into statistical patterns – which does not fit into traditional categories of copying, distribution, or adaptation as it employs aggregation on a massive scale. This novel use of scholarly publishing

highlights that significant economic value can be extracted from free academic content.

Instead, the use and transformation of text could be said to represent a new form of content utilization.

Decker (2025) raises important concerns regarding who benefits from open access policies, and whether beneficiaries should include well-funded AI companies with substantial venture capital backing.

Implications for Academic Attribution

The rise of AI systems trained on open access content also poses risks to academic attribution systems. Unlike human researchers who tend to cite specific sources, AI systems typically aggregate content from several sources, breaking connections between knowledges and their lineages (Decker 2025). Decker (2025) refers to this as “citation laundering,” which describes the misattribution or hiding of original sources through AI generated content.

This raises several concerns both in the immediate and long-term future. For one, if AI-generated content becomes widely cited without proper attribution to original sources, the foundational researchers who built that knowledge are unlikely to receive appropriate credit. This matters in an already problematic system of promotion and tenure that places, at times, undue emphasis on citation counts as evidence of research impact.

The misattribution or suppression of citations also impacts inter- and cross-disciplinary research, where being able to contextualise and connect concepts or ideas to particular knowledge traditions is a crucial aspect of this work. Researchers may also “miss important cross-disciplinary linkages, and which is of particular relevance when it comes to addressing the global challenges that humanity faces,” says Decker (2025).

Potential Solutions and Future Directions

Several approaches are being explored to address these challenges. For instance, API-based access represents one promising solution, says Weinberg (2025). Collections could provide bot-optimized data access points rather than forcing bots to use human-oriented web interfaces. This is something Wikimedia has implemented for the time being—API users receive consistent access to reliably formatted data in exchange for a fee (Weinberg, 2025).

COAR has also established an “**AI Bots and Repositories Task Force**” in July 2025. The initiative represents an important step toward coordinating solutions. According to their **news release**, the Task force will aim to release a report in the fall of 2025 that articulates the problem, documents available mitigation strategies, and includes recommendations for repositories that do not cause problems for legitimate human users attempting to access sites.

The fundamental tension between maintaining open access principles and protecting repository infrastructure requires careful navigation. As Weinberg (2025) observes, institutions may hesitate to restrict access, because “doing so would create barriers for the type of users they hope to invite into their collections.”

Thus, one of the challenges will be to develop solutions that can differentiate between legitimate users and problematic bots without compromising the accessibility that makes open scholarship valuable.

Moving forward, the open scholarship community may need to develop new frameworks and re-articulate existing ones that can account for AI’s role in knowledge creation and dissemination. This may include tweaking existing arguments for alternative impact indicators, ensuring open access policies evolve to address the realities of AI bots and crawlers, and implementing technical solutions that protect infrastructure while preserving access.

This is no small task, but the stakes are high: if current trends continue without effective intervention, the sustainability of open access infrastructure could be compromised—potentially even forcing repositories to implement restrictions and enclosures that undermine the very principles they were designed to uphold.

Responses from INKE Partners

INKE Partner James MacGregor (Director of Research Infrastructure and Development at the Canadian Research Knowledge Network) understands first-hand the risks of this kind of activity:

As caretakers of the technical infrastructure that serves the Canadiana and Héritage collections, collectively consisting of 65M images of Canadian documentary history, we are especially concerned about problematic access to large-scale datasets. One example occurred in 2023 when access to Internet Archive was briefly interrupted globally due to a poorly-behaving bot **attempting to bulk-ingest its OCR data**. (This service interruption pales in comparison, and intent, to the outright malicious cyberattack **suffered by Internet Archive in Fall 2024**.) Internet Archive didn't link the 2023 bot activity outright to an AI organization attempting to harvest its content, but the observed behaviour aligns.

According to James, there are at least two common scenarios that CRKN is concerned about when it comes to AI and crawling/scraping activity:

1. Thanks to the code-generating capabilities provided by genAI, it's now trivially easy to create scripts to crawl a website or other resource, and yet difficult to ensure that the code acts responsibly if said person isn't a coder and doesn't know how or why to write responsible code. A script and a few bucks worth of high-performance compute executed poorly can have a significant impact on a web resource. As highlighted in the Internet Archive blog post, it's *good* practice to proactively declare any sort of possibly-intensive crawling or scraping activity – but it's not *common* practice, especially by non-experts.
2. In order to further train their LLMs, GenAI companies are extremely hungry for any and all data they can access. The usual methods for blocking crawlers – adding rules to robots.txt, or blocking specific user-agent access – rely on the honour system and can be easily ignored or circumvented by bad actors, after which it can be near-impossible to prove that a given LLM has ingested one's content. And once it's ingested, it's impossible to get out.

James points to a new and intriguing proposed extension to robots.txt that aims to fix both the genAI bot *and* licensing problems in one fell swoop: the development of the **“Really Simple Licensing”** or RSL standard and protocol:

The standard aims to coalesce a machine-readable content licensing

approach, similar to what Creative Commons provides, within robots.txt: when a bot visits a site, it is provided within the bots file some machine-readable licensing information that indicates what is permissible with the site contents, and how to obtain that permission. The licensing and compensation terms are tailored to the genAI access concern: access can be free, attribution, pay-per-crawl, or even pay-per-inference. The RSL Collective, the administrative body governing the development of the RSL Standard and associated practices, is also developing automated content licensing mechanisms to support this effort. This is a new approach, and the genAI companies have to date been quiet about it. Time will tell.

James agrees that more work is needed, however:

Since this is still technically the open web, so much depends on companies behaving as good Internet citizens, even with the development of standards like RSL. Infrastructure providers will need to ensure that crawlers respect robots.txt, that user agents are who they say they are, and to ensure that data is accessed responsibly. In the academic digital research infrastructure space, we are fairly behind on this effort.

Resources

Anubis, a configurable open-source firewall: Aery, Sean. 2025. Anubis Pilot Project Report – June 2025. Duke University Libraries. <https://hdl.handle.net/10161/32990>.

Canadian Repositories Community of Practice October Call. October 30, 2025, 1PM – 2PM (ET). – Repositories in the Age of AI: The Attack of the Bots. Register here: <https://www.carl-abrc.ca/mini-site-page/canadian-repositories-community-of-practice-october-call-repositories-in-the-age-of-ai-the-attack-of-the-bots/>

References

Coalition of Open Access Repositories (COAR). 2025. “COAR Launches AI Bots and

Repositories Task Force.” Coalition of Open Access Repositories, July 16. <https://coar-repositories.org/news-updates/coar-launches-ai-bots-and-repositories-task-force/>.

Decker, Stephanie. 2025. “The Open Access – AI Conundrum: Does Free to Read Mean Free to Train? (Guest Post).” *The Scholarly Kitchen* (blog). April 15, 2025. <https://scholarlykitchen.sspnet.org/2025/04/15/guest-post-the-open-access-ai-conundrum-does-free-to-read-mean-free-to-train/>.

Hellman, Eric. 2025. “AI Bots Are Destroying Open Access.” *Go To Hellman*: March 21, 2025. <https://go-to-hellman.blogspot.com/2025/03/ai-bots-are-destroying-open-access.html?m=1>.

Hinchliffe, Lisa Janicke. 2025. “Are AI Bots Knocking Digital Collections Offline? An Interview with Michael Weinberg.” *The Scholarly Kitchen* (blog). June 23, 2025. <https://scholarlykitchen.sspnet.org/2025/06/23/are-ai-bots-knocking-digital-collections-offline/>.

Panitch, Judy. 2025. “Library IT vs. the AI Bots.” *UNC University Libraries*, June 9. <https://library.unc.edu/news/library-it-vs-the-ai-bots/>.

Shearer, Kathleen, and Paul Walk. 2025. “The Impact of AI Bots and Crawlers on Open Repositories: Results of a COAR Survey, April 2025.” Survey. Confederation of Open Access Repositories. <https://coar-repositories.org/news-updates/open-repositories-are-being-profoundly-impacted-by-ai-bots-and-other-crawlers-results-of-a-coar-survey/>.

Weinberg, Michael. 2025. “Are AI Bots Knocking Cultural Heritage Offline?” GLAM-e lab. <https://glamelab.org/products/are-ai-bots-knocking-cultural-heritage-offline/>.

Search



Archives

Select Year



Categories

[Community News](#)

[English](#)

[French](#)

[Insights and Signals Reports](#)

[Observations](#)

[Observations and Responses](#)

[Policies](#)

[Responses](#)

Tags

[AI bots / Robots d'indexation IA](#) [AI governance / Gouvernance de l'IA](#)

[AI safety / Sécurité de l'IA](#) [Berlin Declaration / Déclaration de Berlin](#)

[Bethesda Statement / Déclaration de Bethesda](#) [bibliodiversity / bibliodiversité](#)

[Budapest Statement / Déclaration de Budapest](#) [Canada](#) [Canadiana.org](#)

[Canadian government/le gouvernement du Canada](#) [CAPOS](#) [CARL / ABRC](#)

[collaboration](#) [community engagement / engagement communautaire](#)

[Compute Canada / calcul Canada](#) [copyright / droits d'auteurs](#)

[credibility / crédibilité](#) [CRKN / RCDR](#) [Cybersecurity / Cybersécurité](#)

[data management / gestion des données](#)

[diamond open access / le libre accès diamant](#) [Digital Commons / Commun numérique](#)

[digital scholarship / version numérique](#) [en français / French](#) [English / en anglais](#)

[events and gatherings / événements et rassemblements](#)

[Federation for the HSS / Fédération des sciences humaines](#)

[funding agencies / organismes de financement](#)

[generative artificial intelligence / l'intelligence artificielle générative](#)

[identity management / gestion de l'identité](#) [implementation / mise en oeuvre](#) [INKE](#)

[International OA Week / Semaine internationale du libre accès](#)

[international policy / politique internationale](#)

[licensing agreements / accords de licence](#) [Multilingualism / Multilinguisme](#)

[Naylor Report / le rapport Naylor](#) [open access / libre accès](#)

[open data / données ouvertes](#) [open education / éducation ouverte](#)

[open government / gouvernement ouvert](#) [open infrastructure / infrastructure ouverte](#)

[Open Scholarship Press](#) [open science / science ouverte](#)

[open social scholarship / approches sociales des savoirs ouverts](#)

[open source software / les logiciels libres](#) [ORCID](#) [peer review / critique des pairs](#)

[Perpetual Access / Accès perpétuel](#) [PKP](#) [Plan S](#)

[Plan S update / mise à jour du Plan S](#) [policy / politique](#)

[policy guide / guide des politiques](#) [promotion et titularisation](#) [publishing / édition](#)

[RDC / DRC](#) [RDM](#) [RECODE](#) [recommendations / recommandations](#)

[reports / les rapports](#) [repositories / les dépôts](#)

[research creation / recherche-crédation](#)

[research evaluation / l'évaluation de la recherche](#)

[research libraries / les bibliothèques de recherche](#)

[research output / les résultats de la recherche](#)

[research security / sécurité de la recherche](#) [RPT / révision](#)

[scholarly communication / la communication savante](#) [SFU Library / Bibliothèque](#)

[social media / les médias sociaux](#) [Tri-Agency / des trois organismes](#) [UK](#)



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

