Les robots IA sont-ils en train de remodeler le libre accès ?

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

Read in English

Ce rapport « Insights and Signals » a été rédigé par Brittany Amell, avec ses remerciements au partenaire d'INKE James MacGregor (Réseau canadien de documentation pour la recherche) pour sa relecture et ses commentaires.

Traduction française révisée par Olga Ziminova (MA), Electronic Textual Cultures Lab (ETCL).

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

Résumé

Les robots d'indexation IA modernes envahissent les dépôts en libre accès à travers le monde, obligeant potentiellement les institutions à choisir entre la protection de leur infrastructure et le maintien des principes d'ouverture. Ce rapport « Insights and Signals » présente brièvement les « robots », certains des problèmes qu'ils posent, ainsi que quelques premières réactions de la communauté du libre accès et de la recherche ouverte.

L'évolution du comportement des robots Web

Le paysage du trafic web automatisé a fondamentalement changé au cours des 25 dernières années. Traditionnellement, les robots d'indexation semblaient fonctionner avec un certain code de conduite numérique : ils respectaient les protocoles d'exclusion des robots (ou robots.txt) qui indiquaient les parties d'un site auxquelles ils devaient ou ne devaient pas accéder, s'identifiaient clairement et maintenaient des taux de requêtes raisonnables (Hellman 2025 ; Weinberg 2025). Il arrivait parfois qu'un bot spamme un site avec des inscriptions, mais les administrateurs pouvaient les bloquer assez facilement en se basant sur l'adresse IP. Comme le dit

Hellman (2025), ils faisaient « partie du paysage, sans en être une caractéristique dominante ».

De nos jours, cependant, les bots ne semblent plus être conçus comme avant.

Les bots IA modernes affichent ce que Hellman (2025) décrit comme un comportement « irréfléchi », utilisant au maximum les connexions serveur disponibles et multipliant les requêtes dès que des capacités supplémentaires deviennent disponibles. Contrairement à leurs prédécesseurs, ces bots utilisent souvent des chaînes d'agent utilisateur aléatoires, opèrent à partir de grands blocs d'adresses IP et peuvent se retrouver piégés dans des boucles sans fin, effectuant des milliers de requêtes pour des liens non fonctionnels. À bien des égards, le comportement abusif des robots est impossible à distinguer de celui des attaques par déni de service distribué (DDOS), dans lesquelles l'attaquant utilise intentionnellement des pratiques d'usurpation d'identité et des demandes d'accès automatisées pour submerger un serveur web cible. Cette approche agressive a entraîné une consommation importante de ressources, provoquant le ralentissement de certains systèmes et le plantage complet d'autres.

Défis techniques et opérationnels

L'impact sur les dépôts en libre accès a été considérable. L'enquête menée en 2025 par la Coalition of Open Access Repositories (COAR) a révélé que plus de 90 % des 66 dépôts ayant répondu à l'échelle mondiale ont rencontré des problèmes liés à des robots agressifs, souvent plusieurs fois par semaine, entraînant une dégradation des performances ou des interruptions complètes du service (Shearer et Walk 2025). Ces perturbations obligent le personnel des dépôts, déjà débordé, à se démener pour trouver une solution leur permettant de mettre en œuvre des mesures de protection tout en respectant les principes du libre accès.

Les dépôts ont pour vocation de rendre les connaissances accessibles et utiles, mais lorsque les robots agressifs continuent de causer des problèmes, ils peuvent être contraints de limiter l'accès à leurs ressources.

Le phénomène de « swarming » (essaimage) est devenu particulièrement préoccupant. Comme l'explique Weinberg (2025), il s'agit d'un grand nombre de

robots qui visitent simultanément une collection, téléchargent tout ce qui est disponible et suivent tous les liens détectables. Ce comportement diffère de celui des utilisateurs humains typiques, qui ont tendance à se concentrer sur des domaines de contenu spécifiques.

En essayant de tout récolter sans discernement et en même temps, ces robots créent des pics de trafic imprévisibles et écrasants pour le serveur d'un dépôt.

Mesures défensives et leurs limites

Le personnel des dépôts a tenté d'atténuer et de repousser les attaques des robots, avec plus ou moins de succès (Panitch 2025). Cependant, si certaines de ces mesures peuvent réussir à bloquer les robots, « il est également clair qu'elles entravent l'accès aux dépôts par d'autres acteurs plus bienvenus, tels que les individus humains et les systèmes bienveillants », écrivent Shearer et Walk (2025, p. 1) dans leur rapport pour le COAR.

Par exemple, beaucoup se sont tournés vers des services commerciaux tels que Cloudflare pour bloquer les robots, mais comme le note Hellman (2025), cela a eu des répercussions sur les « bons » robots : « Internet Archive ne peut plus enregistrer les instantanés de l'un des meilleurs éditeurs en libre accès, MIT Press, en raison du blocage de Cloudflare. »

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).

L'efficacité des protocoles traditionnels tels que robots.txt a également diminué. Alors que ce système de conformité volontaire fonctionnait bien avec les générations précédentes de robots d'indexation, de nombreux robots d'entraînement à l'IA ignorent désormais complètement ces demandes, ce qui

représente un écart par rapport aux normes et à l'étiquette établies sur Internet : « Le protocole [robots.txt] ne s'est pas révélé aussi efficace dans le contexte des robots qui créent des ensembles de données d'entraînement à l'IA. Les personnes interrogées ont indiqué que robots.txt est ignoré par de nombreux robots d'IA (mais pas nécessairement tous). Cela a été largement considéré comme une violation des normes d'Internet et un comportement déloyal en ligne » (Weinberg 2025).

Considérations juridiques et relatives aux licences

Le cadre actuel d'octroi de licences pour les contenus en libre accès est également confronté à de nouveaux défis à l'ère de l'IA, qui a entraîné un besoin insatiable de données d'entraînement supplémentaires, écrit Decker dans un **récent article publié sur The Scholarly Kitchen**. La plupart des contenus universitaires en libre accès utilisent des licences Creative Commons, en particulier CC-BY, mais comme le souligne Decker (2025), ces licences ont été conçues en pensant aux lecteurs humains et aux scénarios de réutilisation traditionnels.

L'ampleur considérable et l'appétit pour les données d'entraînement de l'IA ne correspondent pas exactement aux catégories traditionnelles de copie, de distribution ou d'adaptation, explique **Decker (2025)** :

> Lorsque les modèles d'IA ingèrent du texte, celui-ci est désagrégé en « tokens » (c'est-à-dire des mots) qui sont transformés en réseaux neuronaux, lesquels constituent à leur tour la base des réponses aux requêtes. Le texte est converti en modèles statistiques, ce qui ne correspond pas aux catégories traditionnelles de copie, de distribution ou d'adaptation, car cela implique une agrégation à grande échelle. Cette nouvelle utilisation de l'édition scientifique met en évidence la valeur économique considérable qui peut être tirée du contenu académique gratuit.

On pourrait plutôt dire que l'utilisation et la transformation du texte représentent une nouvelle forme d'utilisation du contenu.

Decker (2025) soulève d'importantes questions concernant les bénéficiaires des politiques de libre accès et la question de savoir si ces bénéficiaires devraient inclure les entreprises d'IA bien financées et bénéficiant d'un soutien important en

capital-risque.

Implications pour l’attribution académique

L’essor des systèmes d’IA formés à partir de contenus en libre accès pose également des risques pour les systèmes d’attribution académiques. Contrairement aux chercheurs humains qui ont tendance à citer des sources spécifiques, les systèmes d’IA agrègent généralement des contenus provenant de plusieurs sources, rompant ainsi les liens entre les connaissances et leurs origines (Decker 2025). Decker (2025) qualifie ce phénomène de « blanchiment de citations », qui décrit l’attribution erronée ou la dissimulation des sources originales par le biais de contenus générés par l’IA.

Cela soulève plusieurs préoccupations, tant dans l’immédiat qu’à long terme. D’une part, si le contenu généré par l’IA est largement cité sans attribution appropriée aux sources originales, les chercheurs fondamentaux qui ont construit ces connaissances ont peu de chances de recevoir les crédits qui leur sont dûs. Cela est important dans un système de promotion et de titularisation déjà problématique qui accorde parfois une importance excessive au nombre de citations comme preuve de l’impact de la recherche.

L’attribution erronée ou la suppression de citations a également un impact sur la recherche interdisciplinaire et transdisciplinaire, où la capacité à contextualiser et à relier des concepts ou des idées à des traditions de connaissances particulières est un aspect crucial de ce travail. Les chercheurs peuvent également « passer à côté de liens interdisciplinaires importants, ce qui est particulièrement pertinent lorsqu’il s’agit de relever les défis mondiaux auxquels l’humanité est confrontée », explique Decker (2025).

Solutions potentielles et orientations futures

Plusieurs approches sont à l’étude pour relever ces défis. Par exemple, l’accès basé sur l’API (interface de programmation d’application) représente une solution prometteuse, selon Weinberg (2025). Les collections pourraient fournir des points

d’accès aux données optimisés pour les robots plutôt que de les obliger à utiliser des interfaces web conçues pour les humains. C’est ce que Wikimedia a mis en place pour l’instant : les utilisateurs de l’API bénéficient d’un accès constant à des données formatées de manière fiable en échange d’une redevance (Weinberg, 2025).

En juillet 2025, la COAR a également créé un « groupe de travail sur les robots IA et les dépôts ». Cette initiative représente une étape importante vers la coordination des solutions. Selon le communiqué de presse, le groupe de travail aura pour objectif de publier à l’automne 2025 un rapport qui exposera le problème, documentera les stratégies d’atténuation disponibles et inclura des recommandations pour les dépôts qui ne causent pas de problèmes aux utilisateurs humains légitimes qui tentent d’accéder aux sites.

La tension fondamentale entre le maintien des principes de libre accès et la protection de l’infrastructure des dépôts nécessite une navigation prudente. Comme l’observe Weinberg (2025), les institutions peuvent hésiter à restreindre l’accès, car « cela créerait des obstacles pour le type d’utilisateurs qu’elles souhaitent inviter à consulter leurs collections ».

Ainsi, l’un des défis consistera à développer des solutions capables de faire la distinction entre les utilisateurs légitimes et les robots problématiques sans compromettre l’accessibilité qui fait la valeur de la recherche ouverte.

À l’avenir, la communauté de la recherche ouverte devra peut-être développer de nouveaux cadres et reformuler ceux qui existent déjà afin de tenir compte du rôle de l’IA dans la création et la diffusion des connaissances. Cela pourrait inclure l’ajustement des arguments existants en faveur d’indicateurs d’impact alternatifs, la garantie que les politiques de libre accès évoluent pour tenir compte des réalités des robots et des crawlers IA, et la mise en œuvre de solutions techniques qui protègent l’infrastructure tout en préservant l’accès.

Ce n’est pas une mince affaire, mais les enjeux sont importants : si les tendances actuelles se poursuivent sans intervention efficace, la durabilité de l’infrastructure de libre accèspourrait être compromise, ce qui pourrait même contraindre les dépôts à mettre en place des restrictions et des clôtures qui compromettrait les principes mêmes qu’ils sont censés défendre.

Réponses des partenaires INKE

James MacGregor, partenaire d'INKE (directeur de l'infrastructure et du développement du Réseau canadien de documentation pour la recherche, RCDR), comprend parfaitement les risques liés à ce type d'activité :

> En tant que gardiens de l'infrastructure technique qui dessert les collections Canadiana et Héritage, qui comptent au total 65 millions d'images documentaires sur l'histoire du Canada, nous sommes particulièrement préoccupés par les problèmes d'accès aux ensembles de données à grande échelle. Un exemple s'est produit en 2023, lorsque l'accès à Internet Archive a été brièvement interrompu à l'échelle mondiale en raison d'un bot malveillant qui tentait d'ingérer en masse ses données OCR. (Cette interruption de service est toutefois insignifiante par rapport à la cyberattaque purement malveillante dont Internet Archive a été victime à l'automne 2024.) Internet Archive n'a pas directement lié l'activité du bot de 2023 à une organisation d'IA tentant de récolter son contenu, mais le comportement observé correspond.

Selon James, il existe au moins deux scénarios courants qui préoccupent le RCDR en matière d'IA et d'activités de crawling/scraping :

1. Grâce aux capacités de génération de code offertes par l'IA générative, il est désormais très facile de créer des scripts pour explorer un site web ou une autre ressource, mais il est difficile de garantir que le code agira de manière responsable si la personne en question n'est pas un codeur et ne sait pas comment ni pourquoi écrire un code responsable. Un script et quelques dollars de calcul haute performance mal exécutés peuvent avoir un impact significatif sur une ressource web. Comme le souligne l'article du blog Internet Archive, il est recommandé de déclarer de manière proactive toute activité de crawling ou de scraping potentiellement intensive, mais ce n'est pas une pratique courante, en particulier chez les non-experts.
2. Afin de perfectionner leurs LLM, les entreprises de l'IA générative sont extrêmement avides de toutes les données auxquelles elles peuvent accéder. Les méthodes habituelles pour bloquer les robots d'exploration (ajouter des règles au fichier robots.txt ou bloquer l'accès à certains agents utilisateurs) reposent sur le système de l'honneur et peuvent être facilement ignorées ou contournées par des acteurs malveillants, après quoi il peut être presque impossible de prouver qu'un

> LLM donné a ingéré le contenu d'une personne. Et une fois ingéré, il est impossible de le récupérer.

James souligne une nouvelle extension intéressante proposée pour le fichier robots.txt qui vise à résoudre d'un seul coup les problèmes liés aux robots de l'IA générative et aux licences : le développement de la norme et du protocole « Really Simple Licensing » ou RSL :

> Cette norme vise à regrouper dans le fichier robots.txt une approche de licence de contenu lisible par machine, similaire à celle proposée par Creative Commons : lorsqu'un bot visite un site, le fichier des bots lui fournit des informations de licence lisibles par machine qui indiquent ce qui est autorisé avec le contenu du site et comment obtenir cette autorisation. Les conditions de licence et de rémunération sont adaptées aux préoccupations liées à l'accès à l'IA générative : l'accès peut être gratuit, avec attribution, paiement par crawl, ou même paiement par inférence. Le RSL Collective, l'organisme administratif qui régit le développement de la norme RSL et des pratiques associées, développe également des mécanismes automatisés de licence de contenu pour soutenir cet effort. Il s'agit d'une nouvelle approche, et les entreprises d'IA générative sont restées discrètes à ce sujet jusqu'à présent. L'avenir nous le dira.

James reconnaît toutefois que des efforts supplémentaires sont nécessaires :

> Comme il s'agit toujours techniquement d'un web ouvert, beaucoup dépend du comportement des entreprises en tant que bons citoyens d'Internet, même avec le développement de normes telles que la norme RSL. Les fournisseurs d'infrastructures devront s'assurer que les robots d'indexation respectent le fichier robots.txt, que les agents utilisateurs sont bien ceux qu'ils prétendent être et que l'accès aux données se fait de manière responsable. Dans le domaine des infrastructures de recherche numérique universitaires, nous sommes assez en retard dans cet effort.

Potentiellement intéressant :

Anubis, a configurable open-source firewall (« un pare-feu open source

configurable »): 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/

Références

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 logicels 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-création

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 medias sociaux Tri-Agency / des trois organismes UK

UK / Royaume-Uni UNESCO UVic Libraries Érudit