

Impact des crawlers IA sur les archives ouvertes

Quelles implications pour l'Open Access ?

Maurane Raskinet
Gestionnaire Open Science

Les crawlers dans le paysage du web

CRAWLERS “CLASSIQUES”

- Non invasifs
- Respectueux des protocoles (robots.txt)
- Identifiables
- Faciles à contrôler

CRAWLERS “IA”

- Agressifs
- Multiplient les adresses IP
- Peuvent ignorer les protocoles (robots.txt)
- Affectent les performances des serveurs

Les archives ouvertes : une cible de choix

- Métadonnées structurées et de qualité
- Contenu à grande valeur scientifique
- Facettes de recherche nombreuses
- Moins de barrières techniques
- Volonté d'ouverture



Ampleur du phénomène (sondage COAR*)

*Confederation of Open Access Repositories

- 66 institutions sondées

90 % ont été touchées par des
vagues de crawlers IA



80 % ont subi des interruptions
de service répétées



Avons-nous des solutions ?

Blocage par identifiant (Adresse IP, User-agent)

LIMITES

- Changement d'adresses IP ou utilisation de blocs complets
- Difficile à maintenir
- Risque de bloquer les utilisateurs légitimes
- Impact potentiel sur le référencement des plateformes

AUGMENTATION DE LA CAPACITÉ DES SERVEURS

- Solution temporaire et coûteuse
- Contraire à la sobriété numérique

LIMITER LES REQUÊTES OU LES FACETTES UTILISABLES

- Réduit la pression sur les serveurs mais ne résout pas l'invasion des crawlers

Solutions commerciales et pare-feux

CLOUDFLARE, ANUBIS... (Captcha et honeypots)

Très efficace dans le filtrage, mais...

- Coût parfois élevé
- Peuvent manquer de transparence
- Risquent de bloquer les “bons” bots
Ex : Internet Archive bloqué sur certaines ressources MIT Press



INTERFACE ADAPTÉE À L'INDEXATION DES BOTS

- Interface web épurée
- Facilite la récupération des données
- Soulage les serveurs
- Nécessite des technologies récentes

Enjeux techniques et éthiques

- Protéger nos plateformes
- Préserver leur ouverture
- Empêcher le “blanchiment de citations”
- Créer un cadre éthique



Un défi à relever

DÉVELOPPER DES SOLUTIONS TECHNIQUES ADAPTÉES

- Nécessite une meilleure compréhension de ces crawlers

CRÉER DES LICENCES DE RÉUTILISATION DES DONNÉES LISIBLES PAR LES CRAWLERS

- Repose sur le respect de ces protocoles par les bots

“The open access movement was built on generosity, cooperation, and a belief in the public value of knowledge. But generosity without boundaries invites exploitation”

Eugen Stoica (University of Edinburgh)

Merci

Références

1. Aggressive AI harvesting of digital resources—Community groups—Lyrasis wiki. (s. d.). *LYRASILS Wiki*. Consulté le 18 novembre 2025, à l'adresse <https://wiki.lyrasis.org/display/cmtypg/Aggressive+AI+Harvesting+of+Digital+Resources>
2. Amell, B. (s. d.). Les robots IA sont-ils en train de remodeler le libre accès?. *Open Scholarship Policy Observatory*. Consulté le 16 novembre 2025, à l'adresse <https://ospolicyobservatory.uvic.ca/les-robots-ia-sont-ils-en-train-de-remodeler-le-libre-acces/>
3. Casden, J., Romani, D., Shearer, T., & Campbell, J. (2025). Mitigating aggressive crawler traffic in the age of generative ai: A collaborative approach from the university of north carolina at chapel hill libraries. *The Code4Lib Journal*, 61. <https://journal.code4lib.org/articles/18489>
4. Hellman, E. (2025, mars 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>
5. Shearer, K., & Walk, P. (2025). *The impact of AI bots and crawlers on open repositories: Results of a COAR survey* (p. 9) [Survey]. Confederation of Open Access Repositories.
6. Stoica, E. (2025, juin 12). Protecting university repositories from aggressive web scraping – using database rights to retain control over academic content. *Library & University Collections*. <https://libraryblogs.is.ed.ac.uk/blog/2025/06/12/protecting-university-repositories-from-aggressive-web-scraping-using-database-rights-to-retain-control-over-academic-content/>
7. Trapping misbehaving bots in an AI Labyrinth. (2025, mars 19). *The Cloudflare Blog*. <https://blog.cloudflare.com/ai-labyrinth/>
8. Wikimedia Foundation Annual Plan/2025-2026/Product & Technology OKRs—Meta-Wiki. (s. d.). *Meta-Wiki*. Consulté le 23 novembre 2025, à l'adresse https://meta.wikimedia.org/wiki/Wikimedia_Foundation_Annual_Plan/2025-2026/Product_%26_Technology_OKRs

Remerciements

Un grand merci à mes collègues ORBi (Cécile Dohogne, Myriam Bastin)
et collègues IT (Arnaud Georges, Sylvain Danhieux, Charlotte
Rambeaux et Fabian Smagghe) pour leur relecture et leurs conseils.