

Managing Machine Access to Open Repositories in the Age of Generative AI

Petr Knoth

CORE, The Open University, UK

COAR 2025 Annual Conference

12th May 2025

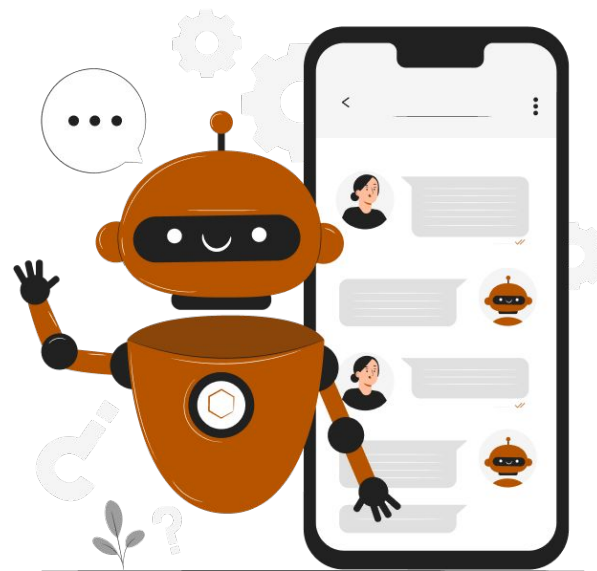


The AI bots problem

The rapid rise of generative AI and its reliance on extensive training data has led to heightened interest in open access repositories, which house legally and freely available as well as trustworthy open knowledge.

While **training** Large Language Models (LLMs) **aligns with Open Access and Open Science principles** and while it provides a promising avenue to mitigate AI's well-documented shortcomings—such as the potential for spreading of misinformation—they have also exposed open repositories to **surging web crawler traffic**.

This phenomenon poses significant challenges, including potential **system overload** and **resource strain** which goes well beyond just open repositories, but puts certain long-established principles of the Web into question.



Context

Makes the argument that: Open access (OA) content, originally intended to be freely accessible to humans, is now being used to train AI models in ways that may conflict with the original purpose of authors, and the scope of open licenses — raising urgent legal, ethical, and policy questions

THE SCHOLARLY
kitchen


ANSEN | APR 16, 2025 | 10 COMMENTS

ABOUTARCHIVESCOLLECTIONS ▾TRANSLATIONS

Guest Post — The Open Access – AI Conundrum: Does Free to Read Mean Free to Train?

It is time for OA proponents to engage in public debate with academic associations, universities and national funding agencies, because the widespread use of academic content in AI models poses significant risks for the research ecosystem.

By **STEPHANIE DECKER** | APR 15, 2025 | **15 COMMENTS**





Context

THE SCHOLARLY
kitchen

ABOUTARCHIVESCOLLECTIONS ▼TRANSLATI

RECENT

Guest Post: Eight Hypotheses Why Librarians Don't Like Retrieval Augmented Generation (RAG)

AI-assisted search is here, and librarians need to have an honest discussion about how to integrate this new technology into library services. This post explores the parallels to the introduction of discovery layers and how to overcome some of the discomfort librarians might have with retrieval-augmented generation.

By **FRAUKE BIRKHOFF** | MAY 8, 2025 | 5 COMMENTS



Makes the argument that: while RAG-based AI tools promise faster and more conversational search experiences (and are capable of **preserving references to original sources**), they introduce significant tensions with core library values, practices, and infrastructures.



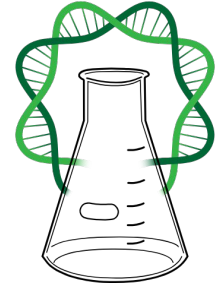
Flawed arguments

Both of the articles use arguments which are fundamentally flawed.



What is wrong with the previous views?

- **Key goals of OA:**
 - Making access to research knowledge public for everyone
 - Accelerating innovation and discovery
 - Public return on investment
- **Copyright** does not protect knowledge embedded in the publications, it only protects the creative expression of that knowledge.



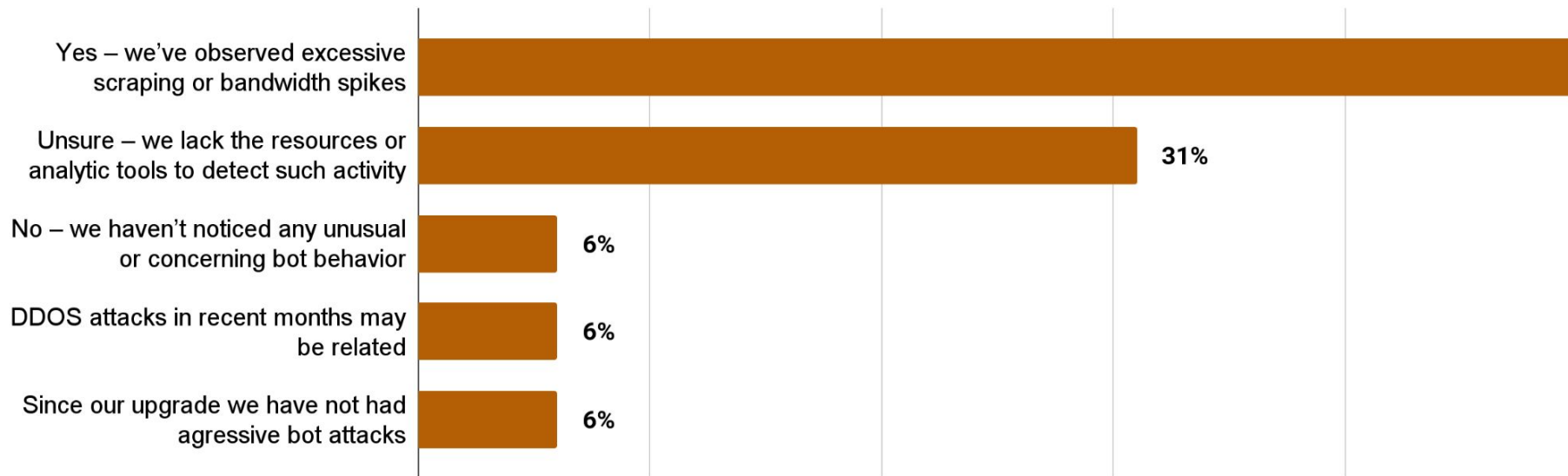
open science

Open science is a global movement that emphasizes the open and transparent sharing of scientific research, data, and knowledge with the broader public



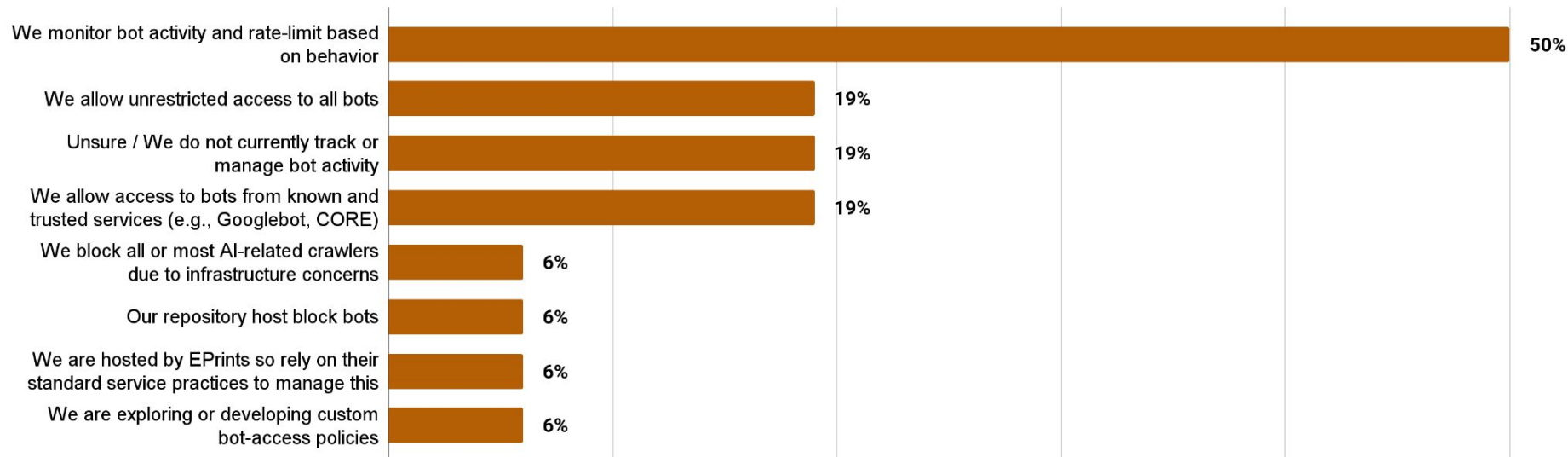
AI Bots survey of CORE Members

Has your institution encountered instances where AI-powered tools or bots have accessed your repository content in unexpected or concerning ways (e.g., excessive scraping, large-scale downloading, or use without attribution)?



AI Bots survey of CORE Members

Which of the following best describes your current or preferred approach to managing machine access (e.g., web crawlers) to your repository?



How is this issue related to making repositories FAIR?

FAIR: Findable, Accessible, Interoperable and Reusable

- Repositories are **not accessible** if bots bring them down
- Repositories are **not interoperable** if they block bots indiscriminately
- If any of the above applies, then it additionally implies that content cannot always be indexed, hence:
 - content in repositories **is not fully findable nor reusable**.



How is this issue related to making repositories FAIR?

*“A **repository** that indiscriminately **blocks machines** from accessing its content **cannot be FAIR**”*



How can access to bots be regulated? 1/2

Mechanism	Description	Enforceability	Blocks Good Bots	Blocks Malicious Bots
robots.txt	A public file that politely requests bots not to access certain pages.	Low	✓	✗
CAPTCHAs	Human verification challenges to block automated access.	Medium	✓	⚠ (often bypassed)
Rate Limiting	Restricts how many requests an IP/user can make in a time window.	High	✓	✓
IP Blocking	Denies access based on IP address or region.	Medium	✓	⚠ (evasion possible)



How can access to bots be regulated? 2/2

Mechanism	Description	Enforceability	Blocks Good Bots	Blocks Malicious Bots
User-Agent Filtering	Blocks or allows traffic based on declared browser identity.	Low	✓	✗
JavaScript Challenges	Requires JS execution, which many bots can't handle.	Medium	✓	⚠
Honeypots	Invisible traps that only bots interact with, revealing themselves.	Medium	✓	✓
Legal Terms (ToS)	Terms of service that prohibit bot access, enforceable via legal action.	Legal-only	✓	✗
Advanced Detection	Uses behavioral analysis and machine learning to detect bots in real time.	High	✓	✓



What can we as the repositories community do to handle the problem?

Option	Description
Manifesto of Good Bot Behaviors	Defining what responsible bot access looks like.
Principles for FAIRbots	Outlining what it means to be a “FAIRbot” and setting clear expectations.
Machine-Readable Bot Protocol	A protocol (like an enhanced <code>robots.txt</code>) to declare repository access policies for bots.
FAIRbot Whitelist	A vetted public list of approved, well-behaving bots for open repositories.
Bot Filtering Service	A filtering and offloading service tailored to repository needs.
Access Log Sharing & Monitoring	Coordinated effort to collect and analyze bot access data across repositories globally.



OR2025 panel on this topic

- What are the acceptable and not acceptable behaviours of machine agents?
- What should best practice guidelines for machine agents accessing repositories look like?
- How to distinguish between machine agents who behave ethically and fairly from highly demanding abusive agents?
- Through which protocols and technical mechanisms can machine agents accessing repositories be informed about the acceptable load they can pose on the target repository system?
- Should there be a community-governed registry of repository bots and what should it look like?



In conclusion...

- The opportunities for the society from the combination of OA and AI are hugely significant.
- Repositories shouldn't block machine access to repositories.
- There is a range of mechanisms using which we can manage this access, but innovation and progress beyond current practice is needed to do this effectively
- The issue has wide implications going far beyond repositories



Discussion ...

