

Open Access Vs. Open Excess: DOAJ And AI Scraper Bots

26/01/2026

*Web scraping to train large language models has surged in the past few years, overwhelming open access infrastructures like DOAJ with massive bot traffic. In this post, DOAJ **Platform Manager Brendan O’Connell** discusses how we’re responding to these challenges.*

Since the launch of ChatGPT in 2022, it’s been impossible to escape endless news headlines about the tremendous promise and peril of AI. And in everyday life, it’s hard not to notice the insertion of AI chatbots into seemingly every digital system we interact with, including search, email, and online shopping.

The rise of AI has been followed with interest by organisations like DOAJ that provide the infrastructure which powers the global open access movement. Over the past year, AI has had significant impacts on our systems and servers, bringing new challenges to open access infrastructures like us.

Large language models (LLMs), such as OpenAI’s GPT (best known for its popular chatbot, ChatGPT) or Google’s Gemini require constant feeding in the form of human-generated content to improve their accuracy, breadth, and depth. This feeding has predominantly taken the form of web scraping – capturing the open (as well as often closed-access) content of the entire Internet through the use of bots. This content is then used to train and improve LLMs.

Automated web scraping is nothing new, and has been the key technology underlying search engines such as Google for over 30 years. However, the current investor-fueled AI startup craze means there are now thousands of well-funded companies developing and deploying their own scraping tools to train AI models, alongside existing major players like OpenAI and Google.

AI scraper bots threaten the open Internet



Luke Conroy and Anne Fehres & AI4Media

Image: <https://betterimagesofai.org> Licensed using <https://creativecommons.org/licenses/by/4.0/>

The impact of this explosion in web scraping for the open Internet, including open access, library, and cultural heritage infrastructures, has been the functional equivalent of another equally old, but more malevolent, automated scripting technique that is turning 30 in 2026: the Denial of Service (DoS) attack. DoS (or Distributed Denial of Service, DDoS) attacks are when a bad actor seeks to deliberately slow down or take down a website by flooding it with a massive volume of automated bot traffic, temporarily overwhelming available server resources.

2025 was the year that this new AI-driven, excessive web scraping to feed LLMs officially broke large parts of the Internet, with open access and cultural heritage organisations such as Wikipedia, the University of Chapel Hill Libraries, and the Directory of Open Access Books publicly documenting slowdowns, downtime, and increased server costs due to massive increases in bot traffic. The term “bot attack” has now become a catchall for purely malicious DoS attacks *and* the newer phenomenon of spikes in traffic from investor-funded AI scraper bots.

How DOAJ is dealing with scraper bots

Since early 2025, DOAJ has seen traffic to our site steadily increasing. The first six months of last year saw 43% more visits to our site than the same period in 2024, and a steady month over month increase.

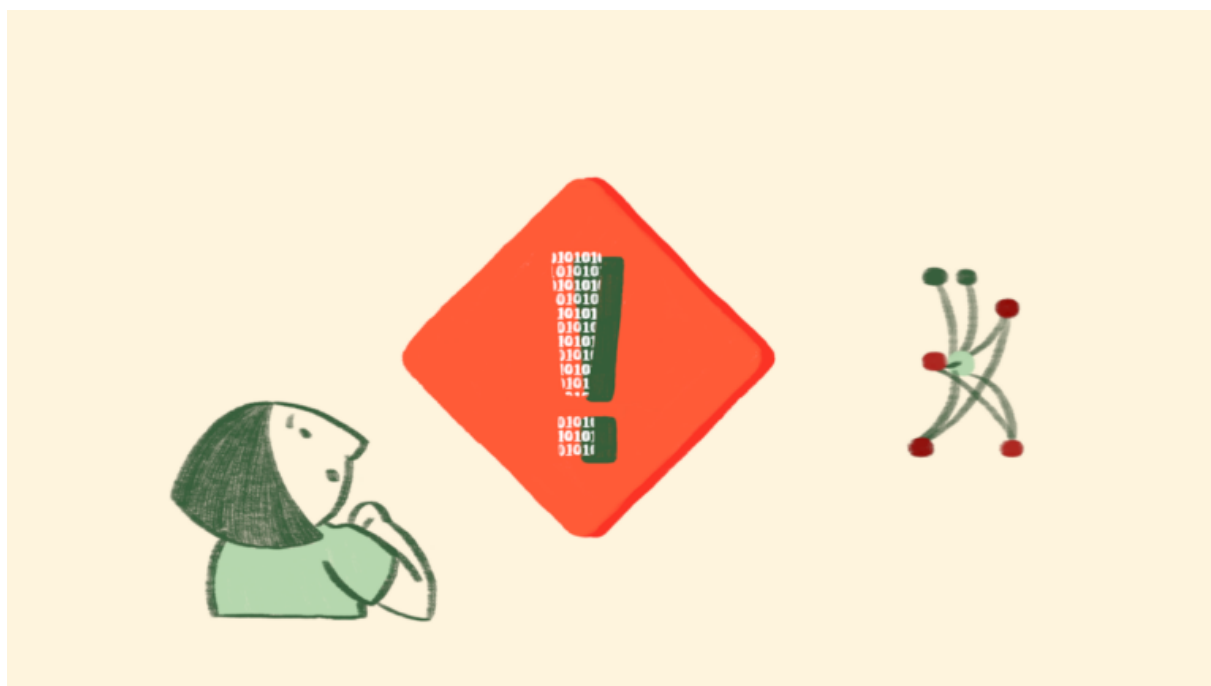
The last six months of 2025 saw a 419% increase over the same period in 2024, culminating in a single day in mid-November where our traffic spiked to 968% greater than a year earlier, resulting in significant slowdowns for users of our public site and our Editorial Team, who use an internal system to evaluate applications from journals for inclusion in DOAJ.

Our system was quickly stabilised in November by adding additional server resources, and we have also implemented selective AI bot blocking and protection strategies from Cloudflare, our cloud services provider. While these quick fixes were effective in stabilising the site, they have increased our server spend significantly.

We are currently implementing a number of changes site-wide to serve static content for user requests, significantly reducing the load on our API with the goal of taking the additional servers offline early this year. For example, our Export Citation feature, which allows users to download article citations, has been re-engineered to asynchronously generate static content in the background. Now, when a human user or scraper bot requests to download a citation from our website, the site first checks to see whether a static version of the content has already been generated through another user request, and serves the static content before querying the API. If the static content doesn't exist, the API is queried and the result is served to the user, which is then saved as static content on our servers for future use.

This change allows us to minimise unnecessary API queries, which are much more computationally expensive than serving static content and thus more prone to create slowdowns for human users when the same endpoints are repeatedly hit by bots. In early 2026, we will be implementing additional changes to our site architecture so that when we face large spikes in bot traffic in future, the administrative workflows of our Editorial team will remain unaffected.

AI creates weird contradictions for OA



Yasmin Dwiputri & Data Hazards Project

Image: <https://betterimagesofai.org> Licensed using <https://creativecommons.org/licenses/by/4.0/>

Open access infrastructures like DOAJ are facing a strange new reality: spending significant staff time and money on blocking *bot* access, in the name of preserving *human* access to open knowledge. This work is further complicated by trying to sort the “good” bots from the “bad” bots: trying to improve how DOAJ articles are displayed in Google search results by *encouraging* Google’s bot crawlers to index us, while *blocking* other scrapers that are slowing down our site for humans.

This work not only costs money and challenges fundamental notions of what it means to be an open access organisation, but as Lyrasis’ Community AI Discussions Working Group notes, “every hour spent addressing AI harvesting is an hour not spent on something else more directly related to the institution’s mission.”

Another strange irony of AI for open infrastructures like DOAJ is that even as we suffer from the negative effects of excessive bot traffic, we and our peers are also exploring the possibilities of using AI in our organisations. A good example is Zenodo's AI-assisted Repository DEposit and Curation (AIRDEC) project, which seeks to create a “joyful, low-burden, cost-effective, and scalable deposit and curation experience” by integrating AI into their systems. I will be exploring the contradictions inherent in the use of AI by open access infrastructures in a future blog post.

Further reading

The Confederation of Open Access Repositories (COAR) has recently published *Dealing With Bots: A COAR Resource for Repository Managers*, a comprehensive guide to the problem space around badly behaved bots and successful mitigation strategies put in place by OA repositories for dealing with them.

Brendan O'Connell is Platform Manager at DOAJ. He has worked in academic libraries and open access infrastructures as a librarian, software engineer, and product manager for over 10 years. His work at DOAJ focuses on building bridges between user needs and technical solutions to advance the global open access movement.
