The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025

June 3, 2025

Kathleen Shearer and Paul Walk

Every day, multiple bots access the repository at all hours 24/7. We estimate performance degradation due to bot activity about once or twice a day, and at least once a week the system crashes entirely requiring an intervention - typically a service restart.

Survey respondent

Executive Summary

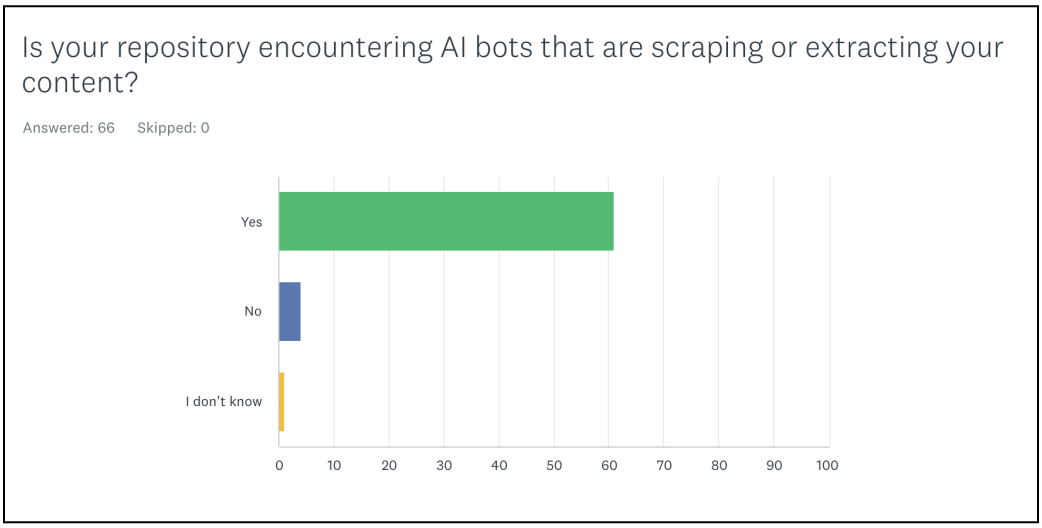
There are a growing number of AI bots crawling repositories. These automated bots, or crawlers, navigate the internet, gathering data and indexing information for search engines, AI and large language models, and other purposes. While some bots are rather innocuous, others are sufficiently aggressive that they are increasingly causing service disruptions in repositories (and other scholarly communications infrastructures). To learn more about the current state and gain a better understanding about the impact of bots and crawlers on repositories, COAR distributed a survey to members in April 2025. The survey received 66 responses from repositories around the world (22 from Canada and US, 22 from Europe, 9 from Latin America, 6 from Asia, 4 from Australasia, 2 from Africa, and 1 unknown).

Over 90% of survey respondents indicated their repository is encountering aggressive bots, usually more than once a week, and often leading to slow-downs and service outages. While there is no way to be 100% certain of the purpose of these bots, the assumption in the community is that they are AI bots gathering data for generative AI training. This type of traffic has shown a marked increase in the last two years or so, and is having a considerable impact on repositories both in terms of the quality of service provision as well as the time and resources required to deal with the bots. In order to mitigate their impact, a variety of measures are being used to minimize or stop AI bots from accessing repositories. Some of the measures being used are considered to be relatively successful in protecting repositories from service disruptions, but it is also clear that they are impeding access to the repositories by other more welcome actors, such as individual human users and benign systems.

The underlying mission of repositories is to provide access to their collections so they are reused and repurposed for the good of scholarship and society. However, the recent rise in aggressive bots activity could potentially result in repositories limiting access to their resources for both human and machine users - leading to a situation where the value of the global repository network is substantially diminished. In order to help the repository community navigate this rapidly evolving landscape and develop solutions that allow repositories to remain as open as possible, COAR will be launching a “Repositories and AI Bots Task Force” in the summer of 2025. The Task Force will bring together technical representatives from repositories and other experts to discuss potential solutions to this problem and develop recommendations for the repository community.

Survey Results

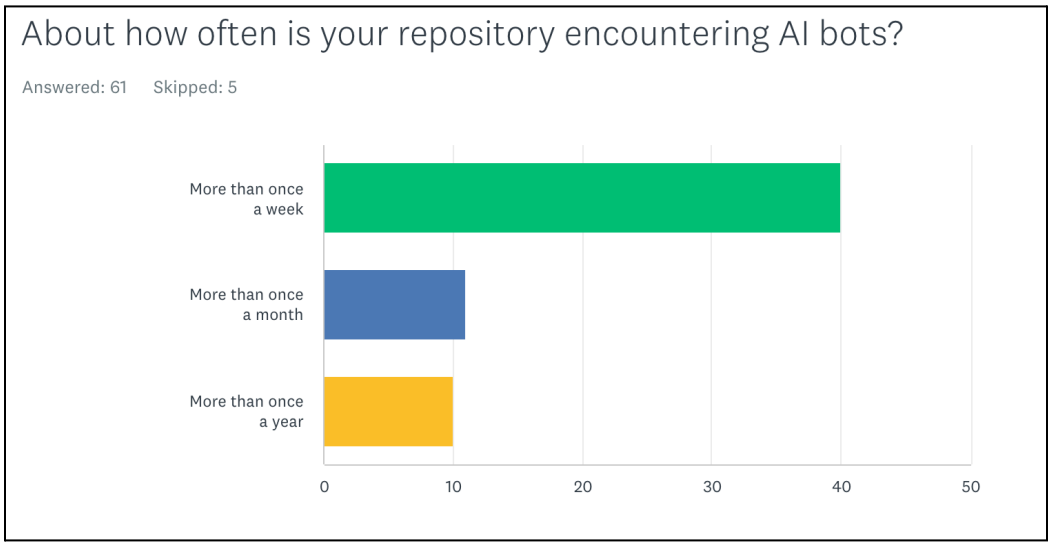
Frequency of AI Bot Encounters



The survey found that almost all repositories are encountering high levels of crawling by bots and anecdotal evidence indicates that there has been a significant increase in these activities over the past couple of years.

No service disruptions, but more than 196,000 accesses in 60 hours.

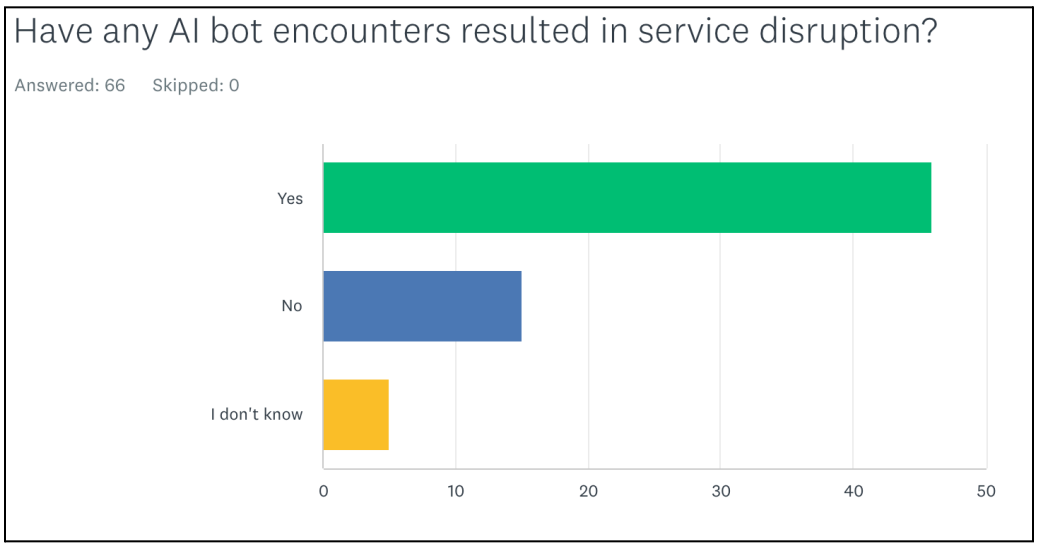
While repositories cannot be sure of the purpose of these crawlers, most assume that this increased traffic is due to bots looking for content to train generative AI models. Several respondents also noted that, in addition to their repository, other library services are also being affected.



Many respondents reported that crawling by bots is happening on a daily basis and nearly constant, but the volume of requests is not consistent. Some repositories have noticed that there are peak times for the bots traffic.

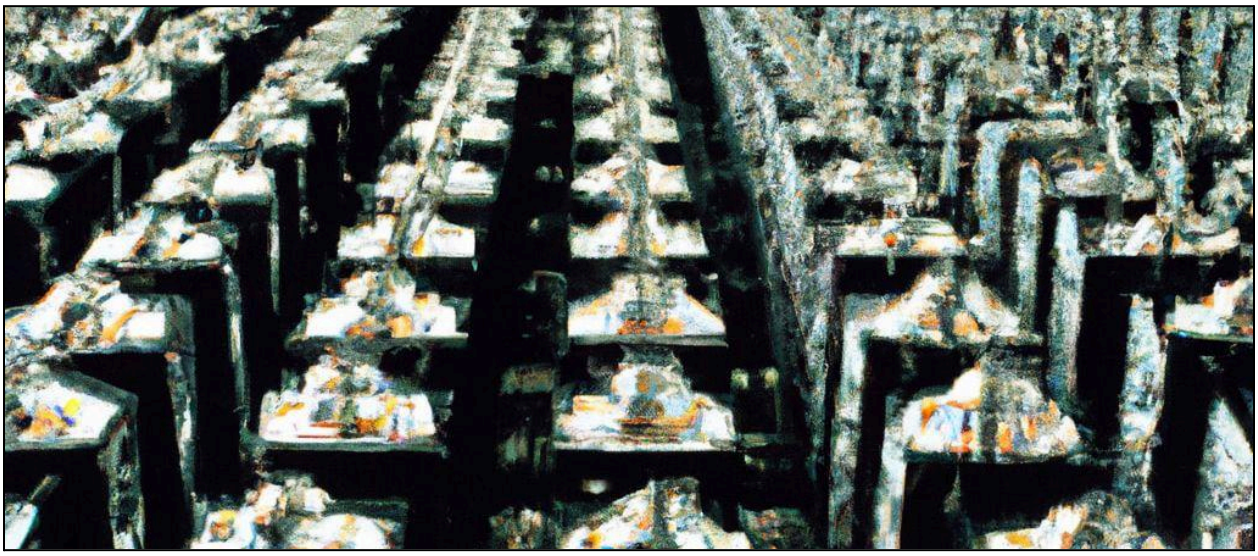
> *A majority of our repository traffic is now ai bots*
>
> *According to our "2024 Bot Usage Activity Report", at least 11 AI Assistants, AI Data Scrapers, and AI Search Crawlers accounted for at least 25% of all bot downloads (503,447 repository downloads) and 15% of all bot pageviews (21,048 pageviews). These AI bots were identified using Dark Visitors - Agent List (https://darkvisitors.com/agents).*

Impact on Repositories



The effects of these bots on repositories are significant. 80% of survey respondents indicated that they had encountered service disruptions as a result of aggressive bots and crawlers. Impacts range from regular service slow downs, to short downtimes, to major service outages sometimes lasting for several days.

About one-third of respondents reported they had experienced at least one down time from 2 to 24 hours; while about one-fifth of respondents said that heavy crawling has led to at least one service outage that has lasted several days. In at least one case, there was a report of long lasting damage to the repository system.



A Generative AI self-portrait by DALL·E. Via Wikimedia Commons

The impact on service interruption depends mainly on how quickly the bots are identified and blocked. If detected promptly, repositories seem to be able to avoid significant down time.

> *It depends on how fast we can put the service up again, in some cases the overcharge takes some parts of the system down. And in some cases we need to block the IPs to avoid the bots that consume the resources.*

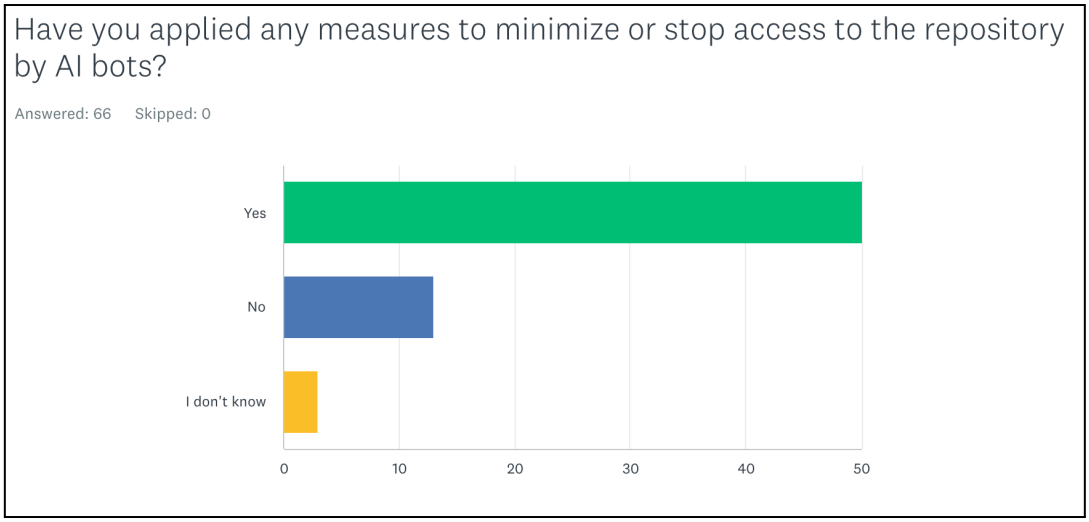
Heavy bot traffic is also having resourcing implications for repositories, in some cases, requiring staff time to resolve service disruptions and applying mitigation measures. On average, repositories are currently devoting about three days a month on this issue, or approximately 10% of a full time employee. Some respondents reported they have only spent a few hours in total in dealing with the impact of bots, but certain repositories, especially larger and more well-established repositories, reported having to devote significant time and resources to this issue.

> *0.5 FTE on this issue per week. Some of the work is education and understanding the issue, some of the work is experimenting, and some of the work is responding to outages.*

> *Just to investigate and block AI, we have 20% of a full time engineer. Our servers are at high temperature because of too much CPU use, so it also has an energy cost.*

Protection Measures

Most respondents indicated that they have applied one or more measures to try to mitigate the deleterious impact of bots.



A variety of approaches are being used, which are listed and described here. It should be noted that these are not always focussed on "AI bots" specifically, but are being applied to any remote software process which might cause the repository's service to be impaired through excessive contact.

Robots.txt

Most repositories implement some version of the "robots.txt" file, and several respondents mentioned this. This file provides a way for any web system (including repositories) to express machine-readable "rules" about which types of external service may access which of their resources. Well-intentioned remote services will be configured to read and respect these rules, but there is no inherent mechanism for enforcing them, and less scrupulous systems may simply ignore them.

One respondent described how, in 2019, they had suffered a service outage due to traffic from Google's official web crawler. They were able to prevent this from recurring by carefully configuring their robots.txt file, and their expectation is that they will be able to do the same for the latest generation of "AI bots". Another respondent mentioned using a service called *Dark Visitors* which provides a publicly curated list of known "artificially intelligent agents" and which can also help to semi-automatically prepare robots.txt files. Other respondents mentioned already configuring robots.txt files to block access by bots from well-known AI services. However, one also commented:

> *For now, those are blocked permanently as a necessary measure to keep our sites stable. These are not sustainable or desirable measures for us or our members.*

Firewalls and IP address blocking

Firewalls are defensive software processes designed to protect systems or, more usually, local networks. As such they are not normally managed within a repository, but more usually by whoever is managing the network infrastructure underpinning the repository system. This may be a host institution, a commercial supplier offering a hosted repository service, or an external network intermediary such as Content Delivery Network (CDN).

> *A firewall works well, but you can still see bots getting through.*

Firewalls work by blocking incoming network connections from external systems, where the latter can be identified by various criteria. A common way to identify external systems is to simply note their IP addresses. In the most basic case, the firewall is informed by a simple list of IP addresses to block.

More sophisticated implementations of this approach use IP address lists which are dynamically generated based on the observed behaviour of remote services. One well-known mechanism mentioned by several respondents is *Fail2ban*, which monitors log files for "suspicious" activity and blocks associatedIP addresses. Other, similar mechanisms mentioned by respondents include components for popular web server software, such as the *mod_evasive* module for *Apache*.

In addition to blocking specific, known IP addresses, some go further by blocking ranges of IP addresses. For example, they may block all IP addresses belonging to a particular "cloud" infrastructure provider. One respondent notes a "... slight risk could block other services". In an even more extreme measure, some respondents reported blocking all IP addresses from entire countries, where those countries were perceived as being the source of a significant volume of unwelcome bot activity. One also reported relaxing such restrictions later, once they were more able to focus their IP addresses restrictions.

> *At times we restricted our sites to national IPs, and received messages from global researchers unable to access our open resources*

Rate limiting

Rate limiting is a counter-measure which is generally configured to set a threshold for the maximum number of requests per second from any given IP address. If this

threshold is exceeded then, usually, additional connection requests are silently dropped and ignored. This measure is normally applied at the level of the repository, rather than the network infrastructure. Again, there are some components which can be added to popular web server software. One respondent mentioned using a system called *Splunk* to provide rate limiting for their DSpace repository.

White-listing "friendly" bots

Although most of the measures described by respondents are based on "blacklisting" undesirable bots - identified either by name or IP address - some respondents have also started to maintain "whitelists" of bots to which they are willing to allow access to the repository. Such lists can be used to complement the other measures.

> *It requires resources on our side to maintain/manage a whitelist.*

Content delivery networks

Content delivery networks (CDNs) are used by several respondents, with *Cloudflare* mentioned in particular. CDNs may provide IP addresses filtering. Furthermore, CDNs such as Cloudflare may add their own measures, such as introducing a CAPTCHA challenge between the repository and remote users or systems. This has the effect of blocking any unrecognised remote system - benign or otherwise - unless that remote system is actively trying to defeat such measures. In other words, such interventions block *all* benign systems, while only blocking *some* (perhaps *most*) malign systems.

Another more recent measure introduced by Cloudflare, mentioned by one respondent, is the *AI Labyrinth*, which recognises when an AI bot has started accessing a website and then uses its own AI process to generate plausible but meaningless content, wasting the time and bandwidth of the crawling system. This approach is also sometimes called a *honeypot*, because it attracts AI bots, where human visitors would more quickly realise that the content was bogus.

Reducing resource-intensive repository functionality

Only one respondent mentioned this approach, of:

> *Disabling resource-intensive functionality like queries, as bot activity on these features are more likely to bring down the site.*

However several respondents mentioned specifically disallowing such resource-intensive pages in their robots.txt files, which is really what robots.txt was designed for.

Effectiveness of Measures

Respondents reported varying degrees of success in terms of reducing the negative impacts of bots with the measures being used. Some respondents felt their methods were relatively successful. But many others said that, while the measures are helping, they are only temporary and more permanent solutions are needed.

> *Together with a significant increase in server resources the implemented measures are currently successful in keeping the system usable for our customers and coworkers. Though each measure was implemented after the last measures became ineffective in dealing with increased / changed scraping traffic.*

> *At present, it's a cat-and-mouse game. While it slows them down, we need to find more proactive and permanent measures. The hunt for a more sophisticated solution is ongoing.*

In terms of impact on users, many respondents indicated that they were aware that they were blocking legitimate users and friendly networks from accessing their repository. This is most likely the case for all the measures mentioned by respondents in this survey.

> *(Measures) are successful in terms of keeping services stable, but they are also highly labor intensive and a blunt instrument. Success would mean allowing legitimate AI harvesting traffic without service degradation.*