glam-e lab

Are AI Bots Knocking Cultural Heritage Offline?

Michael Weinberg

June 2025

v1.0

ENGELBERG CENTER
on Innovation Law & Policy
NYU School of Law

glam-e lab

Table of Contents

1. Introduction 1
2. Methodology 2
    2.1. Key Terms 3
3. Background 4
    3.1. Open Has a History with AI 5
        3.1.1. Recently, AI-Related Concerns Have Expanded to Include Technical Issues 7
    3.2. The Internet Is Full of Bots 8
4. What Have We Learned? 9
    4.1. Every Online Collection Is Technically Unique 9
        4.1.1. There Is No Standard Online Collections System Architecture 10
        4.1.2. Analytics Are Complex, and Not Optimized to Count Bots 10
        4.1.3. The Impact of Bots Is Uneven 11
    4.2. Many, Although Not All, Collections Are Experiencing Disruption from Bots 12
    4.3. AI Training Bots Have Been Operating for Years 15
    4.4. Most Collections Do Not Recognize Bot Traffic Until It Impacts Their Online Presence 15
    4.5. Shared Characteristics of Bots 16
        4.5.1. Some Bots Identify Themselves 16
        4.5.2. Bots Tend to Swarm in Bursts 17
            4.5.2.1. Bursts Tend to Come from Multiple IP Addresses 19
            4.5.2.2. Swarms Exhibit Some Distributed Denial of Service Attack Behavior 20
            4.5.2.3. Swarms Are Increasing in Frequency Over Time 20
        4.5.3. AI Scraping Bots Largely Ignore Robots.txt 20
        4.5.4. Bots Usually Don't Act Like People… 22
        4.5.5. …But Bots May Want the Human Version 22
        4.5.6. Bot Behavior Is Evolving 23
        4.5.7. Non-AI Bots Can Misbehave Too 24
    4.6. Bots Create Problems with Analytics 24
    4.7. Bots Don't Care about Licenses 25
5. Responding to Bots 25
    5.1. Not Everyone Is Relying on AI Bot-Specific Responses 26
    5.2. Simple Fixes Do Not Adequately Reduce Traffic 26
        5.2.1. Updating Robots.txt Has Limited Effect 26
        5.2.2. Reporting Abuse Can Have Some Impact 26
    5.3. Updating Firewall Rules 27
        5.3.1. Blocking by IP Address 27
        5.3.2. Blocking by Geography 27
        5.3.3. Blocking by Domain 27
        5.3.4. Blocking by User Agent String 28

5.4. Increasing Server Capacity and Changing Architecture 28
5.5. Third-Party Bot Countermeasures 29
5.6. Moving Collections Behind Logins 30
5.7. Costs 31
6. What Now? **31**
Appendix A **33**

glam-e lab

About the GLAM-E Lab

The GLAM-E Lab is a joint initiative between the Centre for Science, Culture and the Law at the University of Exeter and the Engelberg Center on Innovation Law & Policy at NYU Law to work with smaller and less-well-resourced UK and US cultural institutions and community organizations to build open access capacity and expertise.

The GLAM-E Lab provides legal counsel to GLAM institutions and cultural organizations as they develop open access programs. The solutions created for those institutions are then integrated into model internal policies and external terms of service that can be adopted by others. The goal of this approach is to use lessons learned from directly representing individual institutions to create self-serve model policies that work "off the shelf" for as many organizations as possible. We supplement these model policies with additional guides and resources to address common challenges.

Acknowledgements

Thanks to Jennie Rose Halperin, Mathilde Pavis, and Andrea Wallace for feedback, guidance, and thoughts. This report would be stronger if it did a better job of integrating their suggestions.

Contact us at info@glamelab.org or https://glamelab.org/

Please cite as: Michael Weinberg, The GLAM-E Lab, "Are AI Bots Knocking Cultural Heritage Offline?" (June 2025) CC BY 4.0.

This project is funded by the UKRI's Arts and Humanities Research Council and the University of Exeter's AHRC Impact Accelerator Account.

1. Introduction

In late 2024, isolated reports began to appear from individual online cultural heritage collections. Those reports described servers and collections straining – and sometimes breaking – under the load of swarming bots. The bots were reportedly scraping all of the data from collections to build datasets to train AI models. This activity was overwhelming the systems designed to keep those collections online.

While concerning, it was not immediately clear if these reports represented a larger trend. Did they reflect the experience of most online collections? Were they outliers? Or early warning signs?

The GLAM-E Lab launched this report to try to answer those questions and start to fill in the bigger picture. It is especially focused on digitized collections and data connected to GLAMs – Galleries, Libraries, Archives, and Museums.

In April 2025, the GLAM-E Lab circulated a short survey to listservs for heritage and technology practitioners in GLAM, both in and outside of academia. That community is international and includes institutions on a range of scales, from flagship national archives to local museums.

We asked institutions with online collections, whether those collections were openly licensed or not, about their experiences with bots building AI training datasets. Those questions focused on the technical impact bots were, or were not, having on their online infrastructure. Our goal was to understand whether the early reports of problems were representative of the wider GLAM community, develop a more nuanced understanding of the technical issues open collections are having with bots, and ultimately begin framing possible solutions.

The initial survey was short, designed to encourage unfiltered sharing. We offered respondents the opportunity to schedule interviews and had more in-depth follow-up conversations with many over Zoom and email.

This report provides a broader snapshot of the intersection of online collections and bots through May of 2025. It is not, and cannot be, comprehensive. Furthermore, the report almost certainly reflects a strong response bias, where individuals experiencing bot-related infrastructure challenges were more likely to respond to the survey and volunteer for follow-up interviews.

Nonetheless, while this report does not capture the definitive experience of the entire GLAM sector's online collections, we believe it does capture the contours of a widespread phenomenon. AI scraper bots are impacting online collections, and their impact is likely to grow.

In brief, we found:

- Bots are widespread, although not universal. Of 43 respondents, 39 had experienced a recent increase in traffic. Twenty-seven of the 39 respondents experiencing an increase in traffic attributed it to AI training data bots, with an additional seven believing that bots could be contributing to the traffic.
- This increase in traffic has been hard to anticipate because few respondents were actively tracking bot traffic prior to the bots triggering a crisis in their collection. Many respondents did not realize they were experiencing a growth in bot traffic until the traffic reached the point where it overwhelmed the service and knocked online collections offline.
- Some respondents have been seeing an increase in bot traffic since 2021, while others did not experience their first spike until 2025.
- Some bots clearly identify themselves, while others take a range of measures to hide their source.
- When bots come, they tend to swarm for relatively brief periods of time. The frequency of these swarms may be increasing.
- Robots.txt is not currently an effective way to prevent bots from overwhelming collections.
- Respondents are deploying a range of home-grown and third-party firewall-based countermeasures to try to screen out bots based on IP address, geography, domain, and user agent string. Some of these efforts appear to be effective, although few are confident that they will be sustainable in the long term.
- Respondents are reluctant to take more aggressive steps to move collections behind things like login screens for a variety of reasons, including concerns about how effective those measures will be in the medium term, that implementing those changes can have negative impacts on welcome users, and whether login-based restrictions run counter to their larger goal of making the collections easily available online.
- Respondents worry that swarms of AI training data bots will create an environment of unsustainably escalating costs for providing online access to collections.

2. Methodology

This document is a response to a series of individual reports in late 2024 and early 2025 of AI data scraping bots impacting digital infrastructures hosting GLAM collections. After seeing these reports, the GLAM-E Lab created a short survey and circulated it among GLAM-focused listservs. The survey questions were intentionally brief, designed to encourage participation and quickly understand how widespread the experience was within the community (see Appendix A).

We received 43 responses from a wide range of institutions in Europe, North America, and Oceania. They ranged in size from large national libraries and museums to smaller community organizations. Respondents included platforms that provide white-label hosting services to a number of institutions.

Once the survey window closed, we reached out to all responding institutions to request a further interview. Some respondents answered questions over email. Others scheduled Zoom calls. Ultimately, we conducted nine realtime interviews, and another eight email interviews. Institutions also shared data, charts, and analytics.

This report includes anonymized data from the survey and follow-up interviews. We have anonymized the data for a number of reasons. First, some institutions were concerned that sharing too much information about their specific practices would tip off the bot operators, undermining the effectiveness of their countermeasures. Second, although their efforts to deploy countermeasures and reinforce infrastructure often illustrated an impressive mastery of limited resources, some participants expressed a level of embarrassment about the quality of resources available to them and the tools they were using to understand what they were experiencing. Third, we felt that anonymity would enable interviewees to share information more freely. As this report is focused on the broader impact within the GLAM sector, we determined that the specific institutional identity tied to any given experience was not particularly relevant to the analysis.

While this report represents more than a compilation of individual stories, it is not a comprehensive study of the entire field's relationship to bots building AI training data. Our goal is to produce a broader snapshot of the current experience in order to anchor ongoing discussions in as much ground truth as possible. As a result, this report contains no conclusions that "all bots are doing this" or "all collections are responding with that." It is a necessarily incomplete description of a set of emerging trends as of the date of publication.

Finally, the GLAM-E Lab's work is primarily focused on making collections open in both the digital sense (available online) and in the legal sense (free from legal barriers to use and reuse, including in commercial context). This report includes data on both open (meaning openly licensed) and online-but-closed collections. In part, this is because the bots do not appear to be altering behavior in response to the presence, or absence, of legal restrictions.

2.1. Key Terms

API: Application programming interface (API) is a way for computers to interact with a website, collection, or other online service. Humans interact with online collections through their web browsers by using things like clickable buttons and graphical interfaces. APIs reduce these interactions to technical interfaces that are more efficiently used by programs and bots.

Bot: Short for "robot," in the context of this paper, bots are programs that move across the internet in search of information. The most important bots in this paper are those tasked with assembling the datasets used to train AI models.

Collection: A set of images, files, text, metadata, and other information related to works hosted by an institution. In the context of this paper, a "collection" is effectively the group of things an institution is making available to the public online, and the infrastructure required to keep it online.

Dataset: A collection of data, including images. This paper is most focused on large datasets used to train AI models. These are usually assembled by the companies developing the models, or by third parties to be used in that development by others.

Distributed Denial of Service (DDoS): DDoS (or just Denial of Service (DoS)) attacks overwhelm servers with traffic. Normally used for malicious purposes such as knocking targeted websites offline or making them unresponsive, they can also occur as unintended side effects of other activity.

Open: Refers to material that is free to view, use, and reuse, including for commercial use and modification.

Scraping: The act of finding, downloading, and formatting data from the internet via automated means.

Server: The computer used to host a collection. A server may be a single computer or a collection of computers and services that make up the technical infrastructure that makes the collection available online.

User agent string: A small snippet of text that software uses to identify itself to servers. The string might identify the browser being used to visit a website (Mozilla/5.0), or the source and purpose of a bot (Googlebot-Image/1.0). While it is usually best practice to use the user agent string to accurately identify the software, its use is optional and not verified.

3. Background

For the past few years, the open movements and communities, like many others, have been wrestling with questions raised by the newest generation of generative AI models. Broadly speaking, these conversations can be broken down into two distinct streams.

The first stream operates at the level of policy, including law and community norms. This often represents existential or philosophical discussions around the nature (and limits) of openness itself. What does it mean when open content created with human users in mind becomes training data for generative AI models? Does the "open" mean open for any use, or just use by humans? Should models that train on open collections have some sort of reciprocal responsibilities to the communities that helped to create the works in those collections, or who make the collections available online? Open licenses typically require

some sort of attribution. Is culturally or legally meaningful attribution even possible for the billions of works incorporated into a single AI training dataset?

The second stream operates at a much more practical, applied level. While it may sometimes feel free for users, the technical infrastructure that makes open possible costs real time and money. It costs money to maintain servers, staff developers, and keep content publicly available. When bots swarm a collection in order to add it to a corpus of training data, they impose costs on whoever is hosting that data in the first place. Is that strain sustainable? Is it reasonable? How much should the cultural institutions that support open movements be expected to pay to make content available to bots scraping them to build AI training datasets?

These conversational streams are deeply related, but also distinct. Problems related to the existential nature of openness require the community to look deep into its collective soul for answers about what it really means to be open. Problems related to paying for increased server costs are less existential, although just as important to the day-to-day availability of open collections. Separating the streams brings rigor to analyzing the challenges, allowing everyone involved to separate concerns that can be addressed with server capacity from concerns that are grounded in fundamental conceptions of openness.

This report focuses on that second, operational stream of discussion. It engages directly with the individuals and institutions that maintain the technical infrastructure that keep collections available online. That includes individual GLAM institutions, as well as the service providers some of them rely on.

3.1. Open Has a History with AI

Online collections tend to be high-quality sources for data to train models. Machine learning and artificial intelligence models (terms that are effectively interchangeable for the purposes of this report) are trained on large quantities of data. These collections are often easily accessible and well structured for machine readability. Many have high-quality metadata that can make information about a given object even more robust.

As such, teams building AI models have drawn on them for some time, regardless of whether they are openly licensed or subject to copyright restrictions. And, in response, the communities behind these collections have a comparatively long history of wrestling with questions related to how, and if, the collections should act as training data for recent generations of AI models.[1]

One of the earliest flashpoints in this relationship was a 2019 controversy over facial recognition models being trained on images posted to Flickr. IBM released a dataset

[1] There is a parallel discussion around how, and if, to integrate the output of generative AI models into online collections. This is no less important to the community. However, it is largely outside of the scope of this report.

consisting of one million faces taken from openly licensed photos on Flickr.[2] Many users who had uploaded images to Flickr responded with surprise.[3]

Most users who uploaded their images to Flickr prior to the release had a specific version of sharing and openness in mind. Very few anticipated the development of facial recognition technologies, or understood how their work might contribute to them. This shift raised new questions. Should their contribution to the commons nonetheless be understood as granting consent for their work (and identity) to be used in this new way? Does "open" mean "open to all," or are there limits? What should it mean when your openly licensed work is used to build technology you find objectionable? Could users have consented to a use they did not anticipate existing when they first shared the image?

Notably, this debate was mostly held at the level of policy – it focused on the meaning and limits of openness, community expectations for it, and possible legal avenues to create boundaries.[4] It included very little discussion of the technical burden imposed on the hosting infrastructure when images were being downloaded to build the training dataset.

Since then, broader conversations about the data used to train AI have expanded significantly.[5] The debate within open movements has tracked that growth, incorporating nuance, new facts, and proposals for paths forward.[6]

These debates are far from resolved. However, even at this stage, we have seen them influence the availability of information well beyond the GLAM sector. A range of publicly available sites have started trying to prevent bots collecting data from visiting, to greater or lesser success.[7] Regardless of the effectiveness of anti-bot efforts on the bots themselves, these efforts may be having negative impacts on non-AI-associated research that uses similar techniques to understand behavior online (albeit at a much smaller scale).[8]

---

[2] John R. Smith, IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems, IBM Blog (Jan. 29, 2019), https://web.archive.org/web/20190313014004/https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/

[3] Olivia Solon, *Facial recognition's 'dirty little secret': Millions of online photos scraped without consent*, NBC News (Mar. 17, 2019), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[4] *See, e.g.* Ryan Merkley, *Use and Fair Use: Statement on shared images in facial recognition AI*, Creative Commons (Mar. 13, 2019), https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/

[5] *See, e.g.* Knowing Machines, https://knowingmachines.org/

[6] *See, e.g.* Anna Tumadóttir, *Questions for Consideration on AI & the Commons*, Creative Commons (July 24, 2024), https://creativecommons.org/2024/07/24/preferencesignals/

[7] Shayne Longpre, et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons* (July 20, 2024), https://arxiv.org/abs/2407.14933

[8] Ryan McGrady, Ethan Zuckerman, & Kevin Zheng, *AI Companies Threaten Independent Social Media Research*, Tech Policy Press (Jan 30, 2025), https://www.techpolicy.press/ai-companies-threaten-independent-social-media-research/

3.1.1. Recently, AI-Related Concerns Have Expanded to Include Technical Issues

In addition to the policy-based concerns that have historically framed these debates, more recently, institutions have raised concerns about the technical impact bots were having on their infrastructure. They reported that bots were coming in large numbers, overwhelming sites. Sites were slowing to a crawl, or being knocked offline entirely, as the result of exponential increases in bot traffic. This is a new, more practical concern about AI training bots: They were breaking infrastructure and running up the costs of hosting digitally available collections.

One of the first public alarms about bot behavior was posted on November 4, 2024, by Bridget Almas to a discussion group hosted by Lyrasis, a nonprofit member organization for library technology.[9] Titled "Aggressive AI Harvesting of Digital Resources," the post opens with a problem statement:

> AI harvesting agents – also known as crawlers, bots, or spiders – are targeting memory institutions . . . The resulting traffic can so impede the service that it is no longer able to function properly, becomes very slow, or goes offline completely.[10]

The post goes on to list behaviors that are commonly shared across the incidents (noting that not all incidents include all listed behaviors):

- the number of simultaneous requests is often very high (up to millions of requests a day)
- requests often come from multiple IP addresses simultaneously. In some cases, over 200 different IP addresses were used by the same harvester to make simultaneous requests
- harvesters sometimes do not follow robots.txt restrictions
- the User-Agent string does not always declare that the user-agent is a bot
- the User-Agent string is often changed for each request, so as to make blocking based on user agent string difficult – it is sometimes hard or impossible to tell harvester traffic from legitimate traffic[11]

As discussed below, these behaviors persist to this day.

This initial post was followed by additional reports. On February 26, 2025, a team member from the Perseus Digital Library, a collection of works from the Greco-Roman world,

---

[9] Bridget Almas, *Aggressive AI Harvesting of Digital Resources*, Lyrasis Wiki (Nov 4, 2024), https://wiki.lyrasis.org/pages/viewpage.action?pageId=364743621

[10] *Id.*

[11] *Id.* The post then listed the ways the behavior differed from traditional DOS attacks, including that the incoming traffic slows when the site goes down, the incidents impact targeted sites over a relatively short period of time, and the behavior does not include active attempts to compromise the site beyond overwhelming it with traffic. These patterns persist today.

announced on Bluesky, "Since last week, we've been experiencing continuing DOS [Denial of Service] from Alibaba subnets that are scraping our content for AI training."[12]

The following month, Wikimedia, probably the largest single source of open content in the world, released a detailed post about its experiences with bots.[13] The post included an explanation of why bot traffic was some of its most expensive traffic to serve. Human readers tend to cluster on similar pages.These popular pages are cached, making them faster and cheaper to serve to users. Less popular pages – at least less popular with people – are served more slowly and more expensively from Wikimedia's core data center. With bots indiscriminately requesting content, therefore increasing demand for that less popular traffic, Wikimedia concluded that 65% of its most expensive traffic was coming from bots.

In June of 2025, the Confederation of Open Access Repositories (COAR) released the results of its own member survey, confirming that survey respondents are regularly encountering bot-related disruption.[14]

These reports were not unique to the GLAM sector. The open source software community was experiencing a similar spike in traffic and undergoing its own debate on how to respond.[15]

Although each of these reports was striking, it was not clear how reflective they were of the broader experience within the digital collections community. Were these stories outliers, idiosyncratic to the collections being targeted? Or did they reflect something that many, if not most, sites were experiencing?

3.2. The Internet Is Full of Bots

In the context of this report, "bots," short for "robots," is a catch-all term for the programs that scour the web, downloading website data to be used later. While there are ways for bots to broadcast the identity of the person or company deploying them, those systems are not mandatory and are not always used. As we will see, that can make tracking the source of bots a challenge.

---

[12] Sarah (@lepidopterane.bsky.social), Bluesky (Feb. 26, 2025, 12:49 PM), https://bsky.app/profile/lepidopterane.bsky.social/post/3lj3wtoa56s2w

[13] Birgit Mueller, Chris Danis, & Giuseppe Lavagetto, *How crawlers impact the operations of the Wikimedia projects*, Diff (Apr. 1, 2025), https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/

[14] Kathleen Shearer & Paul Walk, *The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025*, (June 3, 2025), https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

[15] *See, e.g.* Niccolò Venerandi, *FOSS infrastructure is under attack by AI companies*, LibreNews (Mar. 20, 2025), https://thelibre.news/foss-infrastructure-is-under-attack-by-ai-companies/; Benj Edwards, *Open source devs say AI crawlers dominate traffic, forcing blocks on entire countries*, Ars Technica (Mar. 25, 2025), https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/

Bots are not specific to efforts to build AI training datasets. The behavior – go to a website, process the information it provides, and follow the links out to the next pages – is used in a wide range of contexts. Search engines such as Google use bots to map the internet and incorporate pages into search results. Services such as the Internet Archive use bots to make copies of the entire internet. Researchers use bots to analyze behavior on social networks.

AI training bots are merely one of these bots. Companies building large training datasets – for themselves or clients – send them out across the internet. Each bot goes to a website and downloads everything available – text, images, video, code, etc. They then follow every hyperlink on that website to find other pages, spiraling out in an ever-expanding web of exploration. All of this information is aggregated into large sets of training data. That data is then used to train large AI models.

Individually, AI training bots often behave similarly to other bots. The lines between different types of bots are blurry, and the behavior of an individual bot tasked with building a training dataset may be similar to bots deployed to any number of other purposes.

However, when taken as a group, many site owners view AI training bots differently than other bots. Broadly speaking, there are two ways to think about these differences.

First, site owners distinguish bots based on their perceived utility to the site. A site owner might welcome a Google bot adding it to a search index. They might be less enthusiastic about bots downloading the contents of the site in order to train an AI model for a commercial entity. These differences form part of the policy-based discussion around AI.

Second, site owners distinguish bots based on their traffic demands. A single bot visiting a site once a month will not have a meaningful impact on the site's traffic or hosting infrastructure. In contrast, a swarm of thousands of bots visiting the site simultaneously can quickly overwhelm it and knock it offline. As explored much more deeply below, the impact that high volumes of bot traffic can have on infrastructure forms the core of the technical concerns related to them.

4. What Have We Learned?

The reports of bots impacting online collection are not isolated experiences. This behavior is widespread, and likely growing. There is no single solution to the problems it is creating.

The results of our survey, combined with follow-up interviews, have revealed a more nuanced picture of the impact that bots are having in the community. There are very few universal truths. However, similar patterns do appear across a range of institutions.

4.1. Every Online Collection Is Technically Unique

Although the animating idea of making works available binds them together, each digital collection is built on its own, often idiosyncratic, technical architecture. These

architectures vary wildly in structure and complexity. That can make it hard to develop direct, detailed comparisons between collections. Nonetheless, understanding these differences and broader patterns can help to contextualize some of the broader conclusions below.

4.1.1. There Is No Standard Online Collections System Architecture

Some collections are supported by multi-person, in-house teams of experts who design their own tools. Others outsource their entire operation to third-party services. Many of these third-party services are businesses that exist solely to manage operations for GLAM institutions. Collections within the best-resourced institutions may self-host with effectively unlimited bandwidth capacity and no meaningful constraints on what it costs to serve the data that makes up the collection to any person or bot interested in receiving it. Collections within the smallest institutions may be used primarily to illustrate websites that are mostly optimized to give visitors directions to the physical location. Collections may operate on infrastructure so rickety that it regularly crashed even before bots started showing up, or within a larger system that is unlikely to ever be swamped by traffic to this tiny corner.

The defenses that these collections maintain against the wilds of the internet vary along with their architecture. While the details vary, the broad categories are relatively consistent across platforms.

Firewalls can be used to limit which users can access the system. This allows the teams that run platforms to block individual IP addresses, or blocks of IP addresses, or entire countries. For bots that identify themselves, firewalls can also be used to block bots from specific sources like Facebook or Amazon. Firewall settings are often manipulated by administrators of online collections directly, allowing them to respond to new types of traffic in real time.

Vendors also offer their own specialized anti-bot technology, which often builds upon and enhances other firewalls that collections may already be deploying. The most widely used among respondents and interviewees are the services provided by Cloudflare, although Amazon offers similar services. These rely on a range of techniques to identify bots in a site's traffic and prevent them from overloading a server.

The final option available to a site being overwhelmed by traffic is to simply provision more capacity. If organizations are already hosting their collection with a cloud provider like Amazon Web Services (AWS), this can be as straightforward as a few clicks on a dashboard. It may also involve fundamentally rearchitecting their system and buying new equipment to accommodate the additional traffic. Whatever steps it involves for any given system, the result will almost always include an increase in cost.

4.1.2. Analytics Are Complex, and Not Optimized to Count Bots

Online collections incorporate analytics to capture information about how those collections are used. Until recently, counting bots was not a priority and analytics were not

always deployed to do so. That has made it hard for sites to see the wave of bots coming until it was too late.

In fact, some major analytics platforms screen bots out of their count by default. Traffic numbers in Google Analytics are designed to track how many humans visit the site. Historically, part of the value of those counts was that they screened out bots on the internet because they were not understood as "real" users.[16] Conversely, platforms optimized to protect against bots, such as Cloudflare, provide detailed reports on the number of bots visiting a site.[17]

Interviews with respondents made it clear that each online collection uses its own combination of analytics tools. They may have Google Analytics, or Cloudflare, or both (having recently added one or the other). Alternatively, they may rely on raw server logs and custom scripts to give them insight into user behavior, or a free tier of another analytics platform entirely.

Without specifically deploying analytics optimized to identify bots before they become a problem, it is challenging to track pre-crisis bot growth in a meaningful way. Of course, for collections operating within constrained budgets – which is almost all of them – there was no reason to deploy bot-aware analytics ahead of the problem. As one respondent noted, until recently no one had ever asked them how much of their traffic was bots. The answer did not matter. Another respondent reported that a proposal to increase their ability to monitor bot traffic had been rejected last year because institutional leadership did not believe it was worth the cost to implement.

Yet another respondent explained that the only metric anyone at their institution focused on was the standardized COUNTER[18] measurement of views and downloads. And, in reality, leadership did not care about the COUNTER numbers on their own – only how their institution's numbers compared to those of a historical rival. Another respondent pointed out that the board of their institution was only interested in a top-line total visitor number. As discussed below, this focus on total visitors can present awkward choices once a site realizes that a large portion of their traffic is in the form of bots.

4.1.3. The Impact of Bots Is Uneven

The varying architecture, staffing, institutional priorities, and analytics of online collections means the impact of bots can be uneven. One constant does seem to be that everyone notices when a site goes down, or slows to a crawl, because it is being swarmed by an unsustainably large collection of bots.

---

[16] "In Google Analytics 4 properties, traffic from known bots and spiders is automatically excluded. This ensures that your Analytics data, to the extent possible, does not include events from known bots." *[GA4] Known bot-traffic exclusion*, Google Support (accessed May 22, 2025) https://support.google.com/analytics/answer/9888366?hl=en

[17] *Bot Analytics*, Cloudflare (accessed May 23, 2025) https://developers.cloudflare.com/bots/bot-analytics/

[18] https://www.countermetrics.org/

The impact of bot activity below that threshold is less clear. More traffic is not, in and of itself, a problem. In some contexts, it will not have any marginal impact whatsoever. If a site normally uses 20% of its allocated resources, a sudden jump to 40% utilization for a few hours will not impact the user experience, nor is it likely to meaningfully increase costs.

However, when that 40% jumps again, to 90% or 100%, the impacts are felt immediately. Servers can stop responding, systems can break, and the team behind the site will find itself in an all-hands-on-deck situation.

4.2. Many, Although Not All, Collections Are Experiencing Disruption from Bots

Bots are visiting a large number of online collections, even if not all collections are experiencing an increase in bot traffic. Sometimes, but not always, these bots identify themselves as being deployed by companies building large AI training datasets. For those that do not identify themselves as such, respondents attribute bot purposes based on observed behavior.

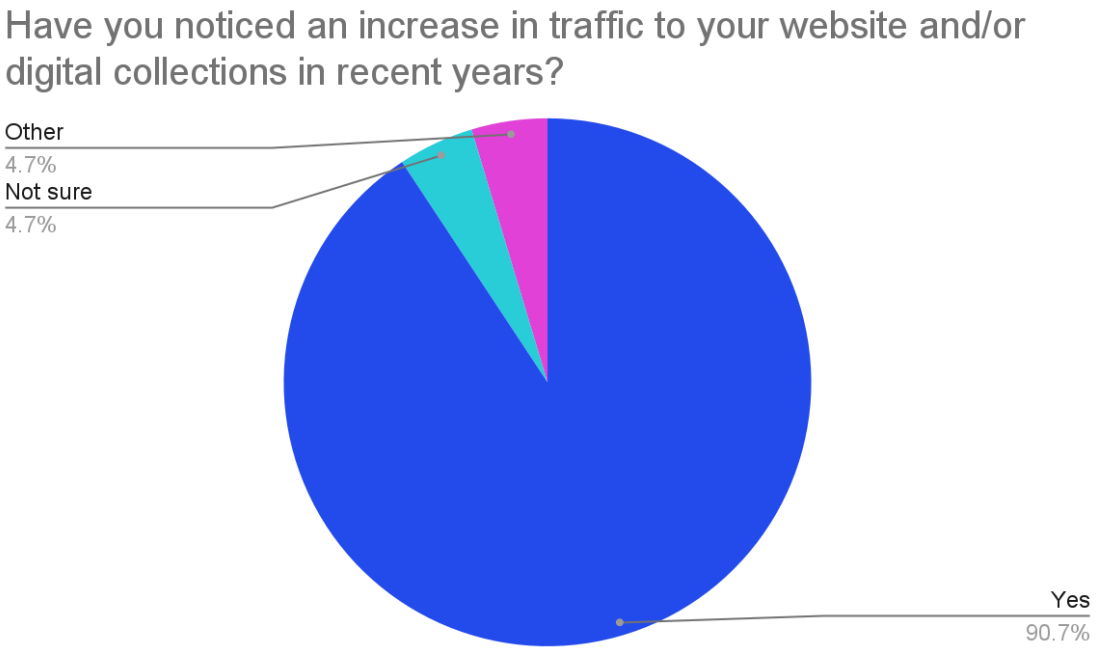

Chart 1: Respondents who have noticed an increase in site traffic in recent years

Forty-three institutions responded to our initial survey. Of those, 39 had recently experienced an increase in traffic. Two were not sure, likely because they did not have analytics in place that would capture such an increase. The remaining two gave unclear responses.

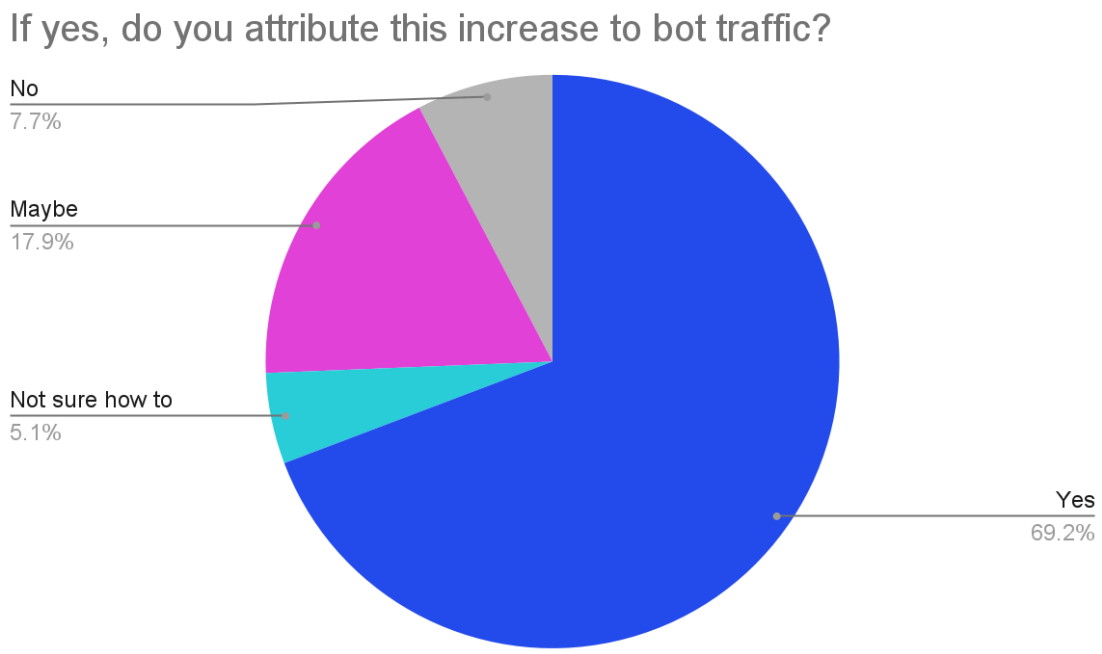

Chart 2: Of respondents who experienced increased traffic, do they attribute traffic to bots?

Of the 39 institutions that indicated they were experiencing an increase in traffic, 27 attributed that increase in whole or in part to bots. Another seven thought that it might be attributable to bots, three did not think it was attributable to bots, and two were not sure how to measure the source of the traffic.

Follow-up interviews suggest that traffic attribution is more art than science. Some institutions utilize services such as Cloudflare that purport to track an increase in bots (although some institutions question the accuracy of those measurements, especially because they are integrated into Cloudflare's bot protection service). Others rely on less explicit measurements, or triangulation through tracking bot behavior. As a result, these responses should be understood as a reflection of how respondents attribute the traffic, not unassailable validation that bots are, in fact, resulting in an increase in traffic.

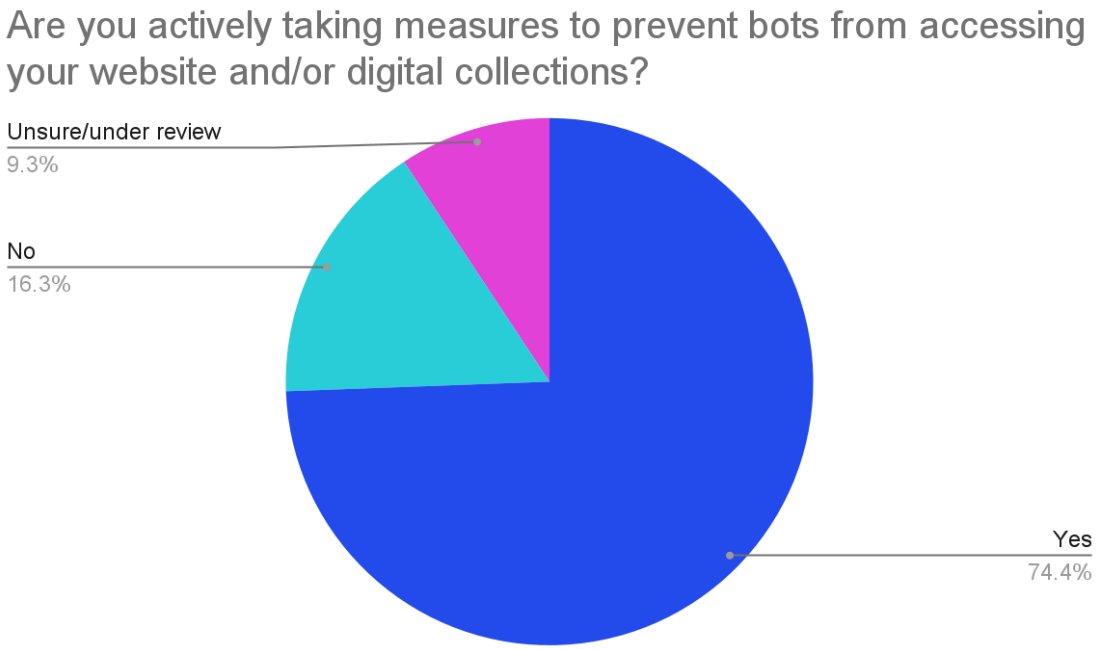

Chart 3: Respondents taking active measures against bots

Thirty-two respondents reported taking active measures to prevent bots. Seven indicated that they were not taking measures at this time, and the final four were either unsure or were currently reviewing potential options.

Again, follow-up interviews suggest that these measures may take a range of forms. Some are extensions of systems the institutions already had in place to protect itself against a broader portfolio of online threats. Others are specially deployed in response to specific experiences with bots. The fact that seven respondents were not currently taking measures to counter bots serves as an important reminder that bots are not universally problematic for collections.

While this survey data cannot provide a comprehensive understanding of how institutions are (and are not) experiencing bots, it does give us confidence that bots are a problem at a scale beyond individual institutions. The initial reports from late 2024 and early 2025 are not isolated incidents. Instead, they are the leading edge of a much broader dynamic washing over online collections.

Bots building training datasets are clearly creating problems for some online collections. However, they are not having a universally negative, or even discernable, impact. Some respondents who had detected bots in their collection said that the bots were not having a meaningful impact on how they support the collection, or the user experience of it. This is a reminder that “Are you seeing bots?” and “Are bots creating problems with your collection?” are distinct questions that do not always have identical answers. It is possible for bots, even bots building AI training datasets, to blend in with the background noise of the internet.

4.3. AI Training Bots Have Been Operating for Years

AI training bots have been accessing collections for a number of years. However, they did not discover all collections simultaneously. Instead, respondents describe a staggered timeline of first contacts between bots and collections.[19]

One respondent could point to a large increase in bot-attributed server load in 2021. Others pointed to points in 2022 or 2023 as the moment when they first began to see significant spikes from these new types of bots. This timeline roughly aligns with the release of public, accessible versions of generative AI models and the rapid rise of public awareness of them.[20]

Other respondents did not detect meaningful bot traffic until much more recently. Some saw their first activity in spring of 2024, and multiple respondents pointed to February or March of 2025 as the first time they had to deal with bots accessing their collections.

Taken together, this suggests that bot activity has been slowly increasing over a number of years. This is likely the result of two separate behaviors. First, existing players engaged in building datasets are expanding the reach of their efforts, discovering new sources of training data over time. Second, new players seeking to build new datasets have spun up new efforts, increasing the likelihood that at least one group of training bots deployed by someone will encounter a given online collection.

4.4. Most Collections Do Not Recognize Bot Traffic Until It Impacts Their Online Presence

As discussed briefly above, prior to the recent growth in bot traffic, most collections had not optimized their analytics to identify (or even count) bots. As a result, many collections were simply unaware that they were being visited by bots until those bots reached a threshold where they significantly degraded the performance of the site.

In practice, this meant that many respondents woke up one morning to an unexpected stream of emails from users that the collection was suddenly, fully offline, or alerts that their servers had been overloaded. For many respondents, especially those that started experiencing bot traffic earlier, this system failure was their first indication that something had changed about the online environment.

---

[19] The reported timeline of first contact between bots and collections may differ from the actual timeline of first contact. This is because, as discussed elsewhere, collections may not have analytics that allow them to recognize early visits from bots.

[20] Dan Milmo, *ChatGPT reaches 100 million users two months after launch*, The Guardian (Feb. 2, 2023), https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Respondents describe scrambling to understand the problem and reset their systems. One respondent explained that the bot traffic was so overwhelming that even their administrative login page was inaccessible.

Others were fortunate enough to see signs of bots in their analytics before the traffic resulted in system degradation. For some respondents, this spike in traffic served as an early warning sign that they needed to prepare for new levels of activity. For others, an occasional moderate spike in traffic remains the only effect the bots are having on the collection – they have not seen a meaningful degradation in collection availability.

The impact of bots on the collections can also be uneven. Sometimes, bot traffic knocks entire collections offline. Other times, it impacts smaller portions of the collection. For example, one respondent's online collection included a semi-private archive that normally received a handful of visitors per day. That archive was discovered by bots and immediately overwhelmed by the traffic, even though other parts of the system were able to handle similar volumes of traffic.

4.5. Shared Characteristics of Bots

Although this report generally refers to bots as an undifferentiated class of programs, they are actually a heterogeneous group of different programs being deployed by different actors for a range of specific purposes. This heterogeneity, combined with the community's limited ability to track and measure their behavior, makes it challenging to make precise observations about what bots, as a group, are and are not doing at any given time.

Nonetheless, there are some trends and behaviors that appear to apply to a significant number of bots across a significant number of platforms.

4.5.1. Some Bots Identify Themselves

Bots can use user agent strings to identify themselves to sites that they visit. This is the equivalent of a bot wearing a "Hello, My Name Is" name tag. For example, a bot associated with Google might identify itself as "Googlebot."[21] This can be helpful in identifying the source of bots visiting specific collections.

[21] *Googlebot*, Google Search Central (accessed May 22, 2025), https://developers.google.com/search/docs/crawling-indexing/googlebot

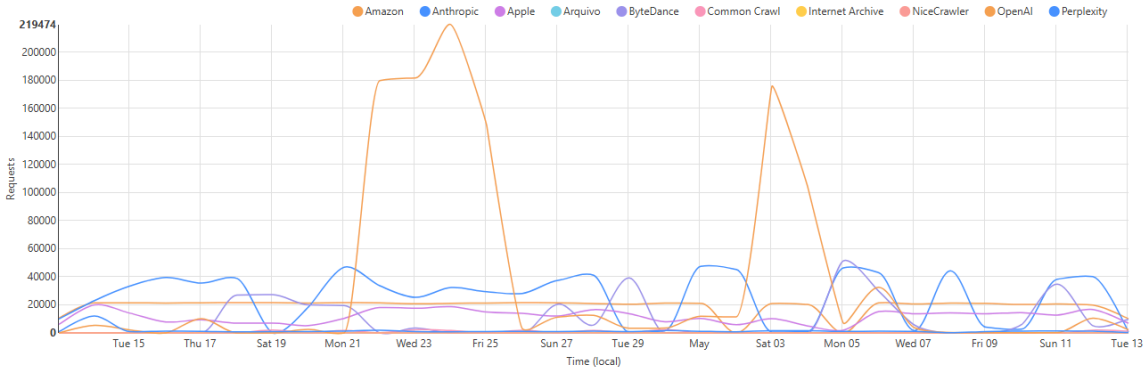

Figure 1: Screenshot of analytics from one of the respondents

Figure 1 is a graph from the analytics dashboard of one of our respondents. The legend at the top of the graph shows the different types of bots visiting the collection over a 30-day period. Properly configured user agent strings can help site owners identify sources of bots and, if appropriate, deploy specific measures against them.

However, just as with name tags at an after-work mixer, there is nothing that requires bots to provide their real identities in the user agent strings. Bots can identify themselves as other bots, or as generic users, or simply leave the string empty.

The general sentiment among respondents was that bigger, more established actors tended to use user agent strings to accurately identify their bots, while lower-profile (not necessarily smaller) actors were less invested in accuracy. While that may be true, there is nothing technically preventing even the most respectable actor from obfuscating the identities of their bots in the user agent string.

4.5.2. Bots Tend to Swarm in Bursts

More traditional bots, such as spider bots indexing the internet for search engines, tend to impose relatively light burdens on servers. These bots visit a site a few times per month, loading pages and moving on.

This is not the pattern exhibited by AI training data bots. Respondents report that AI training data bots swarm in bursts. Both of these behaviors – the swarming of multiple bots, and the fact that the visits came in a compressed period of time – can create problems for online collections.

These bursts often cause servers to slow or fail entirely. Their concentrated nature makes them show up vividly in analytics charts.

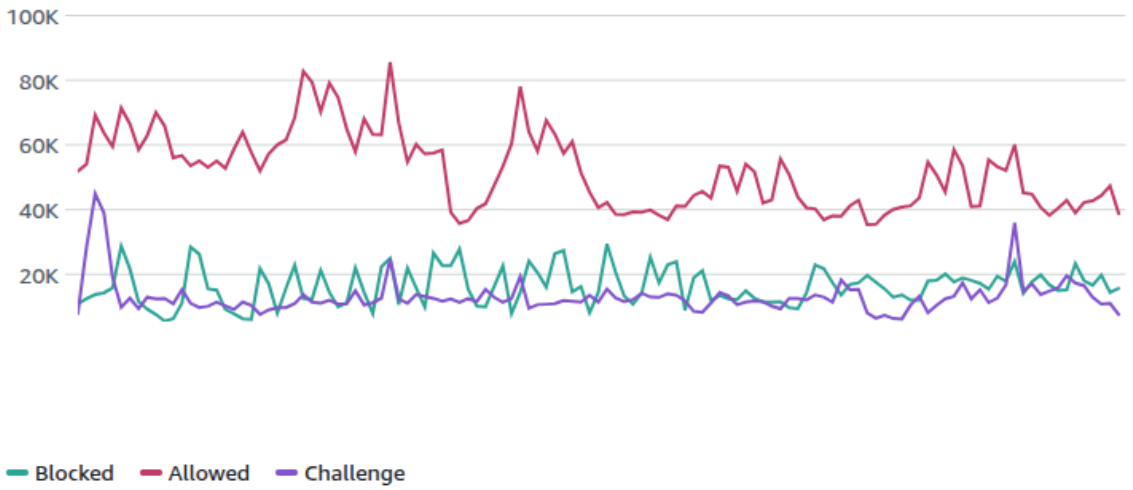

Figure 2: Chart of blocked, allowed, and challenged visitors over a quarter

Figure 2 is a graph from the analytics of one of the respondents. It covers a recent quarter. The respondent is using a service with anti-bot features, and the "blocked" line represents visitors the service believes to be bots. The uneven shape of that line illustrates how bots visit in irregular patterns.

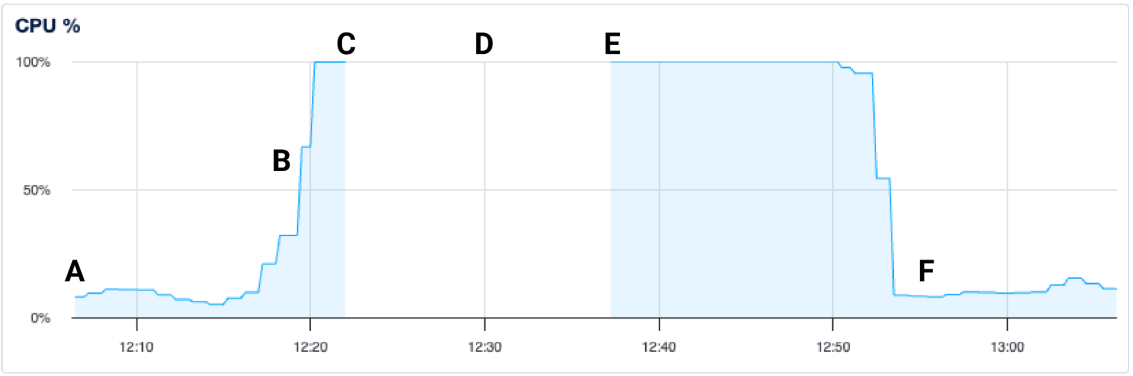

Figure 3: Chart of server CPU load during a bot swarm

Figure 3 is a graph tracking the CPU load of a server during a single hour. At the beginning of the hour (A), the load is hovering at a normal level below 25% utilization. Shortly before 12:20 (B), bot traffic spikes, bringing the load up to an unsustainable 100%. Shortly after 12:20 (C), the team decides to reboot the server. The reboot cycle is represented by a gap in the graph (D). Once the server is back up and running at around 12:37 (E), it immediately returns to 100% load because the bots continue to visit the site. By 12:55 (F), the bots have moved on and the server load returns to pre-swarm levels.

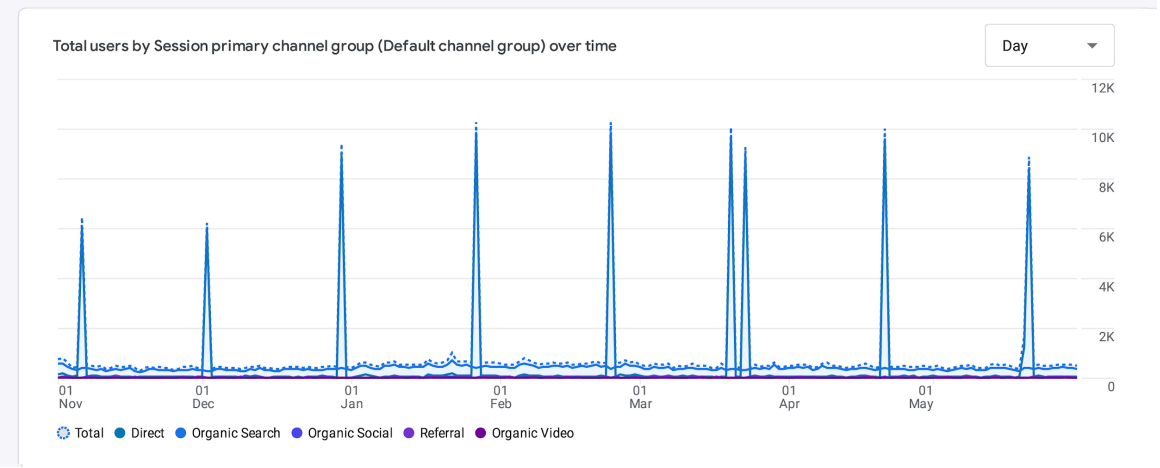

Figure 4: Chart of site visitors over an eight-month period

Figure 4 is an even-more-vivid illustration of bot-related traffic spikes. The collection represented by the graph handles approximately 500 visits on a normal day. About once per month, visits spike to around 10,000. Respondents attribute this spike to bots scraping the collection to build a training dataset.

Online collections design their infrastructure around a set of expectations for anticipated traffic. That means that spikes are always relative. An extra 10,000 visitors in a day to a large collection may not even register on a graph. Alternatively, if those 10,000 visitors represent a 20x increase over expected traffic, it will be enough to knock the collection offline entirely.

4.5.2.1. Bursts Tend to Come from Multiple IP Addresses

Bursts of traffic often come from bots originating from a range of IP addresses. Because blocking the IP addresses of offending bots is one of the ways collections can defend themselves against unwanted visitors, this behavior makes it harder for collections to respond.[22]

Respondents described bots coming from hundreds of thousands, or even millions, of different IP addresses scattered around the world. Individually, these bots are often well behaved and do not exhibit behavior that would create problems. In fact, some may even have been programmed to mimic the behavior of people browsing the collection. However, at scale, these swarms create problems for online collections because of the sheer volume of their requests.

Even if they are spread out across multiple IP addresses, in some instances the bots appear to originate from specific countries or geographic regions. Some respondents were able to identify higher volumes of traffic from smaller, regional ISPs. These ISPs may have been allocated a relatively large number of IP addresses early in the development of the internet, which are now being monetized in service of these types of bot swarms. Other

[22] For more detail about this, *see 5.3.1. Blocking by IP Address*.

respondents reported an increasing number of IP addresses associated with the satellite ISP Starlink.

4.5.2.2. Swarms Exhibit Some Distributed Denial of Service Attack Behavior

Multiple respondents compared the behavior of the swarming bots to more traditional online behavior such as Distributed Denial of Service (DDoS) attacks designed to maliciously drive unsustainable levels of traffic to a server, effectively taking it offline. Like a DDoS incident, the swarms quickly overwhelm the collections, knocking servers offline and forcing administrators to scramble to implement countermeasures. As one respondent noted, "If they wanted us dead, we'd be dead."

However, the creators of these bot swarms do not want to knock collections offline. Disruption to the targeted collection is the result of indifference, not malice. The short bursts of activity may wreak havoc, but once the bots have collected their data, they move on. One respondent estimated that the collection experienced one DDoS-style incident every day that lasted for about three minutes. This was highly disruptive, but not fatal.

4.5.2.3. Swarms Are Increasing in Frequency Over Time

Even short-lived incidents create problems, and an increase in incident frequency has the potential to increase those problems exponentially. Respondents regularly reported that they were seeing an increase in incident frequency over time. Some worried that the activity was moving away from bursts altogether and toward something more sustained. That could create a world where there are more frequent, longer outages for online collections.

This increase in frequency does manifest itself in an overall increase in traffic. One respondent saw its traffic jump from 600,000 visitors in February of 2025 to one million the following month. They attributed that growth to scraping bots. However, as is the case with so much of the reported bot behavior, the increase of swarm frequency is not a universal trend. Another respondent reported a single swarm over a period of five months.

Respondents also raised concern about the timing of the swarms relative to other online behavior. AI scraper bots and search indexing bots arriving simultaneously has caused problems for one respondent. The frequency of search indexing bots has been increasing from one or two per month to almost daily, increasing the likelihood that multiple types of bots arrive at the same time. In such cases, the respondent was not sure how to allocate blame for system failure between the AI scraper bots and the search indexing bots.

4.5.3. AI Scraping Bots Largely Ignore Robots.txt

The problem of unwelcome bot behavior has been solved once. Robots.txt, which uses the Robots Exclusion Protocol,[23] is a tool that allows websites to signal to bots which parts of

[23] *Robots.txt*, mdm web docs (Mar. 13, 2025), https://developer.mozilla.org/en-US/docs/Glossary/Robots.txt

the site the bots should not visit. Its most widely adopted use is to indicate which parts of sites should not be indexed by search engines.

Robots.txt is a voluntary compliance protocol, without the ability to independently prevent bots from visiting any portion of a site. It merely asks, in a standardized, machine-readable way, well-intentioned bots to ignore certain parts of websites. Nonetheless, many bots that operate for a range of purposes are programmed to comply with it. It has served as a "good enough" solution to a number of bot-related problems over the years because bot operators have decided it was in their long-term interest to comply with the requests it communicates.

The 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.

Reports of these types of bots ignoring robots.txt are widespread, even beyond respondents. So widespread, in fact, that there are currently a number of efforts to develop new or updated robots.txt-style protocols to specifically govern AI-related bot behavior online.[24]

These efforts may be enhanced by the EU's Directive on Copyright in the Digital Single Market, which contains provisions designed to create a legally enforceable opt-out mechanism for this type of behavior.[25] However, the language of the directive has not yet manifested into usable mechanisms.[26]

Efforts to update robots.txt, or convince parties deploying bots to be governed by the existing protocol, have not yet provided a practical solution to hosts of online collections. For now, that means respondents are not relying on robots.txt to deter scraper bots.

---

[24] These include specifically extending the Robots Exclusion Protocol to address AI (*see, e.g.* Fabrice Canel and Krishna Madhavan, *Robots Exclusion Protocol Extension to communicate AI preferences vocabulary*, IETF Datatracker (Apr. 4, 2025) https://datatracker.ietf.org/doc/draft-canel-robots-ai-control/, and Paul Keller, *A Vocabulary for Opting Out of AI Training and Other Forms of TDM*, Open Future (Mar. 7, 2025), https://openfuture.eu/wp-content/uploads/2025/03/250307_Vocabulary_for_opting_out_of_AI_training_and_other_forms_of_TDM.pdf), or the creation of a new mechanism to signal preferences with regard to AI (*see, e.g.* Rebecca Ross, *Six Insights on Preference Signals for AI Training*, Creative Commons (Aug. 23, 2024), https://creativecommons.org/2024/08/23/six-insights-on-preference-signals-for-ai-training/)

[25] Article 4, Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market, OJ L 130/92 (17/05/2019).

[26] *See* European Union Intellectual Property Office, *The Development of General Artificial Intelligence from a Copyright Perspective*, at pp. 230-235 (May 2025), https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2025_GenAI_from_copyright_perspective/2025_GenAI_from_copyright_perspective_Full_R_en.pdf

4.5.4. Bots Usually Don't Act Like People…

Some bot swarms distribute their work over a large number of IP addresses and control the behavior of every individual bot in a way that makes it hard to identify them as non-human users. However, there are behaviors that respondents indicated tended to flag a visitor to the collection as automated.

Most humans behave in fairly predictable ways when they land on an online collection page: They will look at the object for six to eight seconds, and then either click a single link on the page or simply move on to some other part of the internet.

Bots behave very differently. When a bot lands on a page, it does not pause to consider the object represented on that page. Instead, it tries to download every version of any images on the page, immediately clicks on every link, and follows those links simultaneously. This behavior flags the user as a bot and massively increases the volume of data being served to it.

This demand is outside of the parameters the system was designed for. Online collections are not designed with the assumption that users would be trying to download every version of every image in the collection on a regular basis.

Respondents report traffic spikes to obscure pages that are suddenly drawing the attention of bots, because bots do not differentiate between pages based on how interesting or relevant the content would be for a human. They are also tracking repeated bot visits to archived files, in a way that would not be logical for a human visitor.

In a blog post, Jason Casden of UNC Libraries described their version of this experience during a swarm in December of 2024: "In November, before we had this problem, we got something like 15 searches with the terms 'Finnish' and 'music.' Basically zero on the scale we operate. On December 4, alone, there were 11,329 searches from thousands of different internet addresses."[27]

This type of unexpected, non-human behavior was one of the key issues Wikimedia discussed in its April 2025 blog post.[28] Wikimedia caches its most popular (for humans) content toward the edge of its network, making it easier and cheaper to deliver. Bots, uninterested in the actual content of the pages, dig deep into Wikimedia's more obscure corners, requesting pages that must be served more expensively from its core datacenter.

---

[27] *Library IT vs. the AI bots*, UNC Libraries (June 9, 2005), https://library.unc.edu/news/library-it-vs-the-ai-bots/

[28] Birgit Mueller, Chris Danis, & Giuseppe Lavagetto, *How crawlers impact the operations of the Wikimedia projects*, Diff (Apr. 1, 2025), https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/

4.5.5. …But Bots May Want the Human Version

One possible solution to the burden that bots impose on online collections could be to provision data specifically for bots via an API.[29] Instead of using a collection's standard, human-optimized, browser-based systems to access content, collections could offer bot-optimized API access points. Wikimedia has successfully used this strategy for some time.[30] In a win for both sides, Wikimedia API users have been willing to pay Wikimedia for access, in return for reliable, and reliably formatted, data.[31] Some respondents indicated that they were exploring similar options.

However, at least one respondent reported considering and rejecting this path. One of the efficiencies with bot-oriented APIs could come from the way the data is formatted. API-served data could remove the page design elements included for the benefits of humans. Instead, it would focus on the data itself, formatted much more efficiently to make it easier for computers to parse.

This respondent was concerned that, even if they offered an API endpoint, the bots would still prefer the version of the site displayed to humans – that those human elements were part of the data the bots collected and used to create the training dataset. If that instinct was correct, the API could end up either fully replicating the human-oriented version of the site (providing no efficiencies) or being ignored in favor of the human-oriented version (providing a waste of resources to build and maintain).

Regardless of the inherent value of the human-oriented elements on a collection item page, any shift to APIs could require those who deploy bots to reengineer their own workflows. Currently, bot managers have constructed workflows optimized to ingest information from human-oriented websites. Partially rebuilding those workflows for each online collection could be burdensome, outweighing any perceived benefit to the bot managers.

4.5.6. Bot Behavior Is Evolving

AI is a rapidly evolving space, and bots scraping the internet for training data are not static. Multiple respondents identified AI-related bot behavior that was expanding beyond traditional scraping for dataset construction.

Some of this behavior has taken the form of search queries that appear to be coming from the AI models themselves. For example, Anthropic recently incorporated web search into its own API.[32] If a user submits a query to Anthropic's Claude AI system, the system can

---

[29] An application programming interface (API) is a way for computers to interact directly with a data source.
[30] https://enterprise.wikimedia.com/
[31] Emma Roth, *Google is paying the Wikimedia Foundation for better access to information*, The Verge (May 22, 2022), https://www.theverge.com/2022/6/22/23178245/google-paying-wikimedia-foundation-information
[32] *Introducing web search on the Anthropic API*, Anthropic Blog (May 7, 2025), https://www.anthropic.com/news/web-search-api

determine that the response would benefit from current information on the internet. In those cases, "Claude generates a targeted search query, retrieves relevant results, analyzes them for key information, and provides a comprehensive answer with citations back to the source material."[33] In other words, Claude itself searches a site and then incorporates those results into its final response to the user. Online collections may start to see bots begin to query their data as part of constructing a user prompt response.

Additionally, some respondents reported an uptick in human (or human-like) visitors referred from AI systems themselves. They speculated that systems such as ChatGPT were including reference links in responses, and users of those systems were clicking on the link to discover the collection. This behavior is similar to more traditional search engine-based user discovery.

4.5.7. Non-AI Bots Can Misbehave Too

While this report is focused on bots deployed to build AI training datasets, respondents regularly mentioned that other types of bots also caused problems. Respondents reported search engine bots sending thousands of requests in minutes, or bots that appeared to be collecting detailed information about scholars at particular universities.

Others returned to the fundamental challenge of distinguishing between AI training data bots and other bots, or of creating a simple heuristic to sort "good" bots that were welcome in the collection and "bad" bots that were causing problems. How distinguishable is a Google bot indexing the site for search from a Google bot sourcing training data for AI models? Is that distinction sustainable as Google integrates generative AI into its search responses?

4.6. Bots Create Problems with Analytics

Many respondents struggle to identify bots with the analytics tools deployed prior to discovering bots on their site.[34] Once they do realize that bots are visiting the collection, it can raise questions about the validity of prior analytics reporting.

The most common of these questions have to do with understanding user growth. Multiple respondents described a dawning realization that traffic growth that had been attributed to an increase in human visitors was, in fact, simply the early signs of a bot swarm.[35] Fully accounting for bots can be especially challenging because many collections are not operating in a static growth environment. They are taking steps to increase human traffic to the collections, and human traffic is actually growing. However, the human traffic growth may not be as dramatic as they had understood for weeks, months, or even years.

---

[33] *Id.*
[34] *See 4.1.2 Analytics are Complex, and Not Optimized to Count Bots.*
[35] A related, although less common, realization was that bots were responsible for skews in platform reporting. Bots tend to identify themselves as coming from a desktop browser. In some cases, bot traffic reversed long-term trends in a collection's analytics away from the desktop and toward mobile.

One respondent described an initial report in March of 2025 that indicated the collection had received one million visits (a significant increase over the historical baseline). After filtering out suspected bot traffic, the team revised that number to 125,000.

Other respondents struggled with how to evaluate the legitimacy of bots. Bots copying the collection into AI training datasets were making use of the collection for a purpose. Should those bots then be considered "users" in their analytics?

Setting aside the philosophical aspects of that type of question, respondents also struggled with the practical implications of answering it. Imagine a collection did view bots building AI training dataset as users, incorporating them into their analytics and future traffic targets. What would happen if the bots evolved and those monthly visitor numbers went down? Are there things they could do to revive bot visitor numbers in order to hit their targets? Would they even want to do them?

4.7. Bots Don't Care about Licenses

There is no evidence from the respondents that bot activity varies between openly licensed collections and those that are merely digitally available. Openly licensed collections do not appear to be more likely to be targeted by bots, and non-open licensing practices do not appear to reduce the likelihood that bots will scrape a given collection.

The legal status of using unlicensed data to train AI models is unresolved in many jurisdictions. If training AI models does not require permission from the data rights holders, the licensing status will not be relevant to entities building training datasets. In practice, even if the licensing status was relevant, it is not always easy for bots to identify the licenses attached to a work it encounters online. As such, bot developers may not attempt to do so.

As a result, there is currently no reason to believe that openly licensing a collection increases the likelihood it will be used to train AI models. Conversely, there is no reason to avoid openly licensing a collection in order to prevent its inclusion in a model training dataset.

5. Responding to Bots

When faced with these challenges, online collections have responded in a variety of ways. The specifics of the responses are influenced by existing technical architectures, staffing levels and expertise, available budgets, and pre-bot expansion roadmaps.

One respondent discussed how their response had been hampered by the internal vocabulary used to describe these types of incidents. The organization's internal security team frames its language in terms of attacks and crisis management. That does not quite capture what the institution is experiencing in this case. The traffic from bots was not a discrete attack or crisis. Instead, it is becoming business as usual. It is not sustainable to

respond to it by repeatedly pressing the flashing “this is a crisis” button. Instead, they were struggling to develop a more stable, long-term plan.

Regardless of specifics when it comes to architectures, budgets, or vocabularies, some bot-response trends did begin to emerge.

5.1. Not Everyone Is Relying on AI Bot-Specific Responses

Some respondents were quick to contextualize their response to bots building AI training datasets across a broader set of threats.

They pointed to a comprehensive set of defenses developed for non-AI-specific threats. These defenses are primarily deployed to counter threats related to vandalism, or the increase in ransomware attacks targeting cultural institutions,[36] or a compromise of data integrity. In some cases, these defenses also provide readily available countermeasures for the types of bots in this report. In others, they serve as a foundation for new defenses that can be quickly deployed and integrated.

Regardless, their utility in responding to AI training data bots illustrates ways in which responses to specific threats need not be specific to those threats. Many broadly applicable security best practices also reduce the negative impact of bot swarms.

5.2. Simple Fixes Do Not Adequately Reduce Traffic

When faced with problematic bot traffic, many respondents first turned to simple countermeasures that had been effective in the past. Unfortunately, these measures have proven to be largely inadequate in these cases.

5.2.1. Updating Robots.txt Has Limited Effect

In recounting their experiences with bots, almost all respondents made at least passing mention of updating their robots.txt file in order to prevent bots from visiting their collections. By and large, these efforts were not successful. While they may have made some marginal impact on traffic levels, no one reported that updating robots.txt has become a significant tool to meaningfully impact traffic levels.

5.2.2. Reporting Abuse Can Have Some Impact

Some respondents have experienced limited success in reporting problematic bots. When bots provided plausible identification via their user agent string, respondents reached out to the parties responsible for the bots to raise concerns. In other cases, when bots were clearly coming from a large third-party infrastructure provider (such as Amazon Web Services), respondents reached out to the third party to report abusive behavior.

---

[36] Zachary Small, *Museum World Hit by Cyberattack on Widely Used Software*, The N.Y. Times (Jan. 3, 2024), https://www.nytimes.com/2024/01/03/arts/design/museum-cyberattack.html

These reports were far from universally effective in reducing traffic. However, they were occasionally effective. They tended to be most effective when third-party platforms were being abused by the parties responsible for the bots.

5.3. Updating Firewall Rules

One of the most widespread, and most effective, set of responses falls under the larger umbrella of updates to firewall rules. Broadly speaking, firewalls can be used to limit traffic from users with certain characteristics. These characteristics can have greater or lesser levels of granularity, and greater or lesser levels of effectiveness.

5.3.1. Blocking by IP Address

Many respondents described taking steps to block bots via IP address, either by individual address or by blocks of addresses. Although not all bots present accurate user agent string information, all of them are associated with an IP address.

Blocking by IP address, especially by groups of IP addresses, can result in overblocking and preventing welcome users from visiting the site. Some respondents attempted to mitigate this by tailoring their blocking strategy to prevent some sets of IP addresses from visiting some subdomains of their collection.

Other respondents reported blocking ranges of IP addresses associated with specific services. For example, one respondent blocked entire IP address ranges associated with Alibaba cloud servers. They determined that the overwhelming majority of traffic from those services was unwanted bot traffic, minimizing the chance that they would inadvertently block welcome traffic.

5.3.2. Blocking by Geography

Respondents also blocked any visitors associated with specific geographies. This approach has obvious downsides, because it also prevents welcome users from those areas from visiting the collection. Nonetheless, if a collection is seeing high levels of bot traffic from regions that are not historically sources of human traffic, the tradeoff may be worth it.

The scope of geographic blocking can vary widely. One respondent noticed a spike in traffic from Dublin. In response, they began blocking any traffic from Dublin without properly configured user agent strings. This additional rule was intended to reduce false positives in the form of blocked human Dubliners.

Other geographic blocking occurred on larger scales. For example, respondents reported blocking all traffic (human and bot) from Singapore, Hong Kong, and Brazil.

5.3.3. Blocking by Domain

For collections not experiencing bot swarms coming from a wide distribution of IP addresses, it can be effective to block by domain. One respondent found success blocking specific domain names such as developers.amazon.com or developers.facebook.com. This likely blocked bots managed by those companies. While this behavior can be effective, it can also provide a perverse incentive for bot managers to hide their identities in order to avoid being subject to domain-level countermeasures.

5.3.4. Blocking by User Agent String

Similarly, some collections have implemented blocking by user agent string. This response can be effective for "well-behaved" bots that properly identify themselves. However, it creates the same negative incentives as domain-level blocking, pushing bot managers to obfuscate their true identities.

5.4. Increasing Server Capacity and Changing Architecture

Firewall rules help collections reduce the amount of traffic to their systems. Increasing server capacity and changing technical architecture address the problem from the opposite direction by increasing capacity to handle that traffic.

These changes can take many forms, like increasing the number of servers or being more aggressive about load balancing and caching. It can also include improving the analytics deployed to monitor the system as a whole.

Expanding capacity costs money, and some collections are better positioned to make these changes than others. Some respondents have experienced minimal bot disruption because the increased traffic came at the end of a recent upgrade cycle.

One respondent explained that they had spent the past few years overhauling their technical infrastructure in order to shift it to a more cloud-based architecture. They had made this change for general operational reasons because it gave them more flexibility in managing their systems. While the overhaul had not been done with AI bots in mind, the new architecture's flexibility allowed them to respond quickly when they started seeing a spike in traffic.

Other respondents described responding to bots by accelerating already planned work. Like the shift to a cloud-based architecture, this work had not been planned with bots specifically in mind. Nonetheless, the general improvements also enhanced the collection's ability to respond to bot traffic spikes.

When it comes to future improvements, respondents indicated that bot traffic and countermeasures will become part of their planning. However, that does not necessarily mean they will be accelerating that planning because of bot traffic.

While many respondents understood that increasing capacity and changing architecture could help them mitigate the impact of bot traffic, they were not necessarily enthusiastic about it. One observed that giving bots a better user experience did not feel like the best use of limited resources. Another noted that leaders are not usually excited to hear that, instead of building something new, the technology team wants to spend time rebuilding something they already have.

5.5. Third-Party Bot Countermeasures

Many respondents reported deploying, or increasing the use of, third-party services that offered the capacity to counter bots.

Of these services, Cloudflare appeared to be the most popular and widely used. Cloudflare offers users the ability to track bots by originating entity (Google bots vs. Anthropic bots vs. Apple bots, etc.) and purports to provide more accurate analytics of bot activity more broadly.

Respondents described noticeable improvements after increasing their use of Cloudflare. One noted that, although they can still see the bot traffic spikes in their Cloudflare dashboard, since implementing protections, none of those spikes had managed to negatively impact the system. Others appreciated the effectiveness of Cloudflare but worried that an environment of persistent bot traffic would mean they would have to rely on Cloudflare in perpetuity.

Assessments of Cloudflare's true effectiveness varied significantly among respondents. Some worried that they did not have a reliable way to independently verify how well Cloudflare was actually performing, or how accurately the analytics provided by Cloudflare reflected the true nature of the operating environment. They also worried that a monoculture of Cloudflare usage made it a target for industrial-scale countermeasures.

Inevitably, respondents experienced user complaints related to overblocking as a result of false positives. A school reached out to one collection because it had been improperly identified as a bot and blocked from access. That collection's Cloudflare implementation did not allow it to whitelist an individual IP address, so there was nothing the collection could do except to tell the school to reach out to Cloudflare directly. That respondent believed that the reports of false positives that actually came to their attention represented a small (but unknown) percentage of actual problems.

Cloudflare is not the only third-party option discussed by respondents. Many used Amazon's AWS to host their collections and had deployed Amazon's bot countermeasures. One respondent noted that this created its own set of questions. When faced with bots running on AWS, the collection deployed AWS bot countermeasures. In effect, they observed, they were paying Amazon to deal with Amazon.

Cloudflare and Amazon are countermeasures used by collections that manage their own technical infrastructure. Other respondents relied on managed collection hosting for third

parties. For these collections, their approach to the increase in bot traffic was somewhat simpler: let the company providing the collection management worry about it (for a price).

5.6. Moving Collections Behind Logins

For both technical and legal reasons, bots tend to be optimized to collect data from publicly available websites. As a result, moving collections behind login screens can significantly reduce the amount of bot traffic they receive.

Although this option was widely understood, respondents tended to be wary of actually deploying it, for a number of reasons.

Some wondered if such a move would actually be effective for long enough to justify the work required to implement the change. Moving collections behind login screens (and associated terms and conditions) could also marginally expand the legal remedies available to use against bots, but few viewed this as a problem with a legal solution.

Furthermore, it is unclear how effectively login screens act as a technical barrier to bots, both today and in the future.[37] At the same time, anti-bot challenges such as CAPTCHAs are increasingly difficult for humans to solve. Respondents worried that, at some point, they would be imposing unreasonable burdens on the people they want to visit their collections in the name of restricting the bots they do not.

However, the larger objection to moving works behind a login screen was philosophical. Respondents expressed concern that moving work behind a login screen, even if creating an account was free, ran counter to their collection's mission to make their collections broadly available online. Their goal was to create an accessible collection, and adding barriers made that collection less available.

A related concern was the impact that any anti-bot measures would have on the bots that were welcome. A respondent explained that many partners build on their collection using a range of automated systems, accessing it in ways that were not designed to work with logins or CAPTCHAs. On some level, a big part of their mission was to allow bots to access the site.

Even many bots that do not work in close coordination with platforms are welcome. Many respondents rely on Google search to help users discover their works. Those respondents actively encourage Google's search indexing bots. Respondents worried that there was not a straightforward way, either technically or conceptually, to quickly categorize bots as "good" or "bad."

[37] Kyle Orland, *AI bots now beat 100% of those traffic-image CAPTCHAs*, Ars Technica (Sep. 27, 2024), https://arstechnica.com/ai/2024/09/ai-defeats-traffic-image-captcha-in-another-triumph-of-machine-over-man/

Login screens and other barriers were being most actively explored in relation to what respondents viewed as sensitive collections. Conversations related to some of those collections, like artifacts from Indigenous communities, build on existing debates within the community.[38] Others, such as a new collection of handwritten historical documents that had recently been transcribed, might be sensitive because of their higher-than-average interest to teams building datasets to train handwriting recognition AI models. Going forward, new collections might be evaluated for bot attractiveness to determine if they should be on the open internet or behind a login.

5.7. Costs

Collections are experiencing a range of costs related to bots. That includes costs in staff time as they respond to swarms, money as new services and servers are deployed, and reputation when sites are down or are working poorly. One respondent explained that, even though their stakeholders were not concerned about these types of operational costs, they might worry about the environmental cost of the increased server traffic.

While real, the costs do not necessarily directly correspond to the size, frequency, or intensity of bot swarms. Some collections are part of sprawling consortia, where spikes in traffic that are large for the collection are insignificant from the larger institutional perspective. Others already had countermeasures such as Cloudflare deployed on their systems, so the primary marginal cost of responding was activating features they were already paying for as part of their subscription.

Nonetheless, many collections are experiencing more direct costs. Deploying new servers can have an unanticipated impact on budgets, and almost all respondents reported devoting more staff time to wrestling with these issues – whether responding to incidents, planning responses, or simply trying to improve analytics to track them. One respondent that acts as a hosted solution for collections explained they had already increased the hosting fees they charge customers by 10% to compensate for the increased traffic.

6. What Now?

Bot traffic that can reasonably be attributed to the construction of AI training datasets is having a real impact on online collections. It is hard to know how broad that impact is, and the specifics of the impact vary according to the collection. Nonetheless, all signs are that the impact is significant and widespread.

That impact comes in the form of costs to the collections, and users unable to access them when servers are slow or offline. It is also causing some institutions to rethink their relationship to online access. What does it mean to be a “user” of an online collection? Are some uses, or ways of accessing collections, categorically beyond the rules? Realistically, attempts to weigh answers to these questions are also being influenced by the incredible valuations of the companies responsible for this latest round of bots. Would answers

[38] *See, e.g.* the work of ENRICH https://www.enrich-hub.org/

change if the growth of AI was not being driven by a small set of for-profit corporations valued at billions of dollars?

Considering the differences in geographies, budget, institutional size, and collections focus, the responses described in this report are surprisingly uniform. This illustrates how much the digital GLAM community operates on a shared set of global norms.

Laws governing scraping and collecting text and data, including images, can vary significantly across jurisdictions, as do social and cultural expectations. The resources available to respondents building and supporting online collections ranged from tightly constrained to effectively unlimited. However, regardless of where they were located, to a first approximation, all respondents approached this problem with the same set of assumptions and a shared set of values. There were differences in the responses, as detailed in this report. Nonetheless, those differences did not obviously track jurisdictional boundaries. This is likely the result of the more-than-two-decades effort to build a global culture around online access. People involved in the community see themselves as a community sharing a core set of values, regardless of their national legal or cultural environment.

The multitude of conversations around revising robots.txt may represent the best path forward for both the online access community and the entities deploying bots to build AI training datasets. The implementation details are complex, but robots.txt has proven surprisingly effective at governing related issues online for decades. That is a reason to be optimistic about its prospects here.

Some of that optimism for finding a workable balance is grounded in a concern that the current path is not sustainable. The cultural institutions that host online collections are not resourced to continue adding more servers, deploying more sophisticated firewalls, and hiring more operations engineers in perpetuity. That means it is in the long-term interest of the entities swarming them with bots to find a sustainable way to access the data they are so hungry for. Will that long-term interest spur action before the online collections collapse under the weight of increased traffic? There is no reason that it must, but there is hope that it might.

Responsible entities building training datasets might even find an advantage in agreeing to support sustainable rules of the road and help build technical measures to enforce them. If complying with those rules gave them privileged access to collections, it might give them advantages over fly-by-night upstarts unwilling, or unable, to follow them.

Perhaps the best hope for the future of online access is that AI training dataset bots fade into the background noise of the internet. They add traffic, but at manageable levels. They are building commercial products off the Commons, but so are plenty of other companies. And maybe they will help more people discover these collections, build on them, and create something new.

glam-e lab

Appendix A

Survey Title: Digital Collections and AI-Related Bot Activity

Survey Introduction: The GLAM-E Lab is conducting a study on how (and if) traffic to digital collections has been impacted by bots building AI training datasets. We are interested in institutions that have and have not noticed this phenomenon, or aren't sure how to recognize it. Once you have completed the form below, we will reach out to discuss next steps. Questions? Email info@glamelab.org

Survey Questions:

1. Name (short answer)
2. Email (short answer)
3. Institutional affiliation (institutions will be anonymized in the final report) (short answer)
4. Have you noticed an increase in traffic to your website and/or digital collections in recent years (yes/no/I am not sure how to measure or detect this/other)
5. If yes, do you attribute this increase to bot traffic? (yes/no/maybe/I am not sure how to measure or detect this)
6. Are you actively taking measures to prevent bots from accessing your website and/or digital collections? (yes/no/other)
7. Would you be willing to share traffic data as part of this study (data would be presented anonymously) (yes/no/maybe)
8. Can we contact you with follow-up questions? (yes/no)
9. Anything else we should know? (long answer)