

DOI:10.1145/3775063

James Grimmelmann

Law and Technology

AI Scraping and the Open Web

Websites turned to the legal system when technical measures against scrapers failed.

GENERATIVE AI COMPANIES and websites are locked in a bitter struggle over automated scraping. The AI companies are increasingly aggressive about downloading pages for use as training data; the websites are increasingly aggressive about blocking them. Their disputes are starting to undermine the open web. Right now, if I write a post for my blog, anyone in the world can read it, and anyone in the world can develop a program to analyze it. That is good for me, good for readers, and good for developers. But when AI scrapers overload servers with their requests, and when websites put up barriers to keep AI scrapers out, the open Web suffers, and we all lose.

A Brief History of Scraping

Tussles between websites and scrapers are not new. Almost since there has been a Web to scrape, people have been scraping it and using the data to make search engines, caches and archives, analytics platforms, research datasets, and more. And for almost as long, some websites have objected and tried to stop

the scraping with a mix of technical and legal measures.

Broadly speaking, scrapers cause two kinds of problems for websites. First, they create bad traffic: millions of automated requests that no human will ever see. Bad traffic drives up websites' bandwidth and hosting costs. In the worst cases, poorly behaved scrapers can cause sites to fail under the load, in a kind of unintentional denial-of-service (DoS) attack.

Second, scrapers can deprive websites of good traffic: the human pageviews that allow them to pay the bills and keep the lights on. Most blatantly, spammers and phishers sometimes create exact duplicates of websites, trying to trick users into visiting them instead. But something similar happens on a smaller scale when a search engine responds to a user's query with a snippet from a page, answering a user's question without the clickthrough.

Websites have used self-help technological approaches to detect automated bot traffic and block it, but these measures have never been fully effective. An HTTP GET request looks the same to a website whether it is coming from a hu-

man or a bot. Sites have tried identifying IP addresses making large volumes of requests, but in response scrapers have learned to spread out their requests, and so on back and forth in a never-ending cat-and-mouse game.

When technical measures fail or do not seem worth trying, websites have turned to the legal system. Some of them have tried to limit access to their sites to argue that the act of scraping itself is illegal under contract law, tort law, or computer-misuse law, such as the federal Computer Fraud and Abuse Act (CFAA). Others have tried to assert ownership over the contents of their sites to argue that downstream uses of scraped data amount to copyright infringement.

Strikingly, all of these different bodies of law have ended up in a similar place. On the one hand, scraping as such is mostly allowed. It is not a tort or a crime to load a website that is available to anyone on the Web, even if you use a bot to do it and the website has told you not to. On the other hand, scraping is not anything-goes. Scrapers are not allowed to crash servers by overloading them with requests, and con-



tent does not drop out of the copyright system just because it was scraped from a public webpage.

The result is that for approximately two decades—from the mid-2000s to the mid-2020s—scraping law has enforced a kind of compromise between websites and scrapers. The visible face of the compromise is the Robots Exclusion Protocol (REP).^a REP itself is extremely simple. Site owners use a file—the famous `robots.txt`—to specify whether to allow or disallow a particular bot to access a particular resource. Crawlers are expected to consult the `robots.txt` file before loading pages from a site, and to refrain from loading disallowed pages.

This compromise is consistent with the handshake deal on which the open Web rests. The terms of that deal are that people who visit a website give enough back to it for the site to continue to exist. Most often, users pay in form of their attention, which websites convert to money by showing ads. Traditional crawlers—such as search engines and

archives—are consistent with this deal, as the law recognizes.

The Rise of AI

Generative AI has upended this rough compromise. Cutting-edge models are trained on as much high-quality data as AI companies can find. The open Web—where websites put billions of pages on view for anyone to read—is an irresistible source.

In scraping the Web for training data, generative-AI companies have hammered websites with bad automated traffic. The increased load might be manageable if there were only a few AI crawlers. But immense investor interest in generative AI means there are thousands of AI startups, many of which are running their own scrapers. A quarter or more of all Web traffic now consists of automated AI crawler requests.

Even worse, not all AI crawlers are well-behaved. Some of them hammer a site all at once rather than spreading requests out over time; others load the same pages over and over again rather than keeping their own caches. Considerate crawler design has always been a best practice rather than a legal

requirement, but now many AI crawlers fall well short of the standard.

At the same time, AI companies are increasingly diverting good human traffic away from websites to their own services. Some of this is an acceleration of the general cutthroat competition for attention that drives the modern Internet. OpenAI's Sora video generator, for example, is playing in the same short-form video space as TikTok and Instagram Reels.

But in other cases, websites can point to specific user visits that AI crawlers deprived them of. Google's AI Overviews, which now appear at the top of most searches, typically provide directly the kinds of information users would previously have clicked through to an underlying website for.

Perplexity's retrieval-augmented-generation approach is even more direct. In response to a user query, Perplexity often goes and fetches a relevant page directly from the Web, then provides the user with an AI-generated summary. Although it provides a link to the summarized page, many users will stop there, never clicking through. From the site's perspective, Perplexity

^a RFC 9309: Robots Exclusion Protocol;
<https://www.rfc-editor.org/rfc/rfc9309.html>

has converted a good pageview (seen by a human, including ads) into a bad one (seen only by a bot).

A New Arms Race

In response, websites have also started backing away from the compromise that characterizes the open Web. Some of them are moving content behind log-in barriers or paywalls. Others are expanding their use of robots.txt to exclude known AI crawlers or to ban bots altogether.

These moves have real downsides for the public. For one thing, they make it harder for human users to browse the Web; people who value their privacy or can't afford to pay are shut off from pages they previously enjoyed. For another thing, this withdrawal shrinks the online commons. Information is harder to find, harder to link to, and harder to archive. Well-behaved bots that provide valuable services like reverse image search or historical archiving are caught in the crossfire. The Internet as a whole becomes less usable and less useful.

Worse still, AI crawlers and websites have started engaging in a technical arms race. Some crawlers respond to robots.txt limitations by changing their user-agent strings, so they can argue that a restriction on "AICompanyBot" does not apply to "AICoBot." Others ignore robots.txt entirely and use completely forged bot names. Some AI companies have even been accused of circumventing paywalls by creating fake accounts.

For their part, websites and their service providers have used increasingly aggressive technical measures to detect and block AI crawlers. Cloudflare, which has visibility into automated access to all of the sites it serves, has been particularly active in offering ways for its customers to block AI crawlers. In parallel, it has developed an "AI Labyrinth" to trap AI crawlers in an infinite maze of dynamically generated pages.

A New Scraping Protocol?

In response to these tussles, various groups have started trying to create new versions of robots.txt for the AI age. Many of these proposals focus on making REP more granular. Instead of just a binary decision—allow or disallow access—they add mechanisms for websites to place conditions on the usage

AI companies are increasingly diverting good human traffic away from websites to their own services.

of the contents scraped from it. This is not the first such attempt—a group of publishers proposed a system called the Automated Content Access Protocol in 2006 that was never widely adopted—but these new ones have more industry support and momentum.

Cloudflare's Content Signals Policy (CSL) extends robots.txt with new syntax to differentiate using scraped content for search engines, AI model training, and AI inference.^b A group of publishers and content platforms has backed a more complicated set of extensions called Really Simple Licensing (RSL) that also includes restrictions on allowed users (for example, personal versus commercial versus educational) and countries or regions (for example, the U.S. but not the EU).^c And Creative Commons (disclosure: I am a member of its Board of Directors) is exploring a set of "preference signals" that would allow reuse of scraped content under certain conditions (for example, that any AI-generated outputs provide appropriate attribution of the source of data).^d

At the same time, some of these same groups are trying to extend REP into something more ambitious: a framework for websites and scrapers to negotiate payment and content licensing terms. Cloudflare is experimenting with using the HTTP response code 402 PAYMENT REQUIRED to direct scrapers into a "pay per crawl" system.^e RSL, for its part, includes detailed provisions for publishers to specify commercial licensing terms; for example, they might re-


quire scrapers to pay a specified fee per AI output made based on the content.

Going even further, other extensions to RSL include protocols for crawlers to authenticate themselves, and for sites to provide trusted crawlers with access to encrypted content. This is a full-fledged copyright licensing scheme built on the foundation—or perhaps on the ruins—of REP.

Preserving the Best of the Open Web

CSP, RSL, and similar proposals are a meaningful improvement on the ongoing struggle between websites and AI companies. They could greatly reduce the technical burdens of rampant scraping, and they could resolve many disputes through licensing rather than litigation. A future where AI companies and authors agree on payment for training data is better than a future where the AI companies just grab everything they can and the authors respond only by suing.

But at the same time, RSL and similar proposals move away from something beautiful about REP: its commitment to the open Web. The world of robots.txt was one where it was simply expected, as a matter of course, that people would put content on webpages and share it freely with the world. The legal system protected websites against egregious abuses—like denial-of-service attacks, or wholesale piracy—but it treated ordinary scraping as mostly harmless.

A world in which that is no longer true is smaller and sadder. More content is hidden away behind paywalls, available only to those who can afford it. Authors' voices carry less far. It is more difficult to create innovative new applications that need extensive data—indeed, the academic experiments that led to today's generative-AI models might not have been feasible if all the content they used were tightly controlled and licensed. RSL keeps open access as an option, but much less so than REP. A truce is better than open warfare, but it would be better still for the Web to be free and at peace. 

b Content Signals; <https://bit.ly/43kzcBe>

c RSL: Really Simple Licensing; <https://bit.ly/49NcntM>

d From Human Content to Machine Data: Introducing CC Signals; <https://bit.ly/4qRPgEq>

e Introducing Pay per Crawl; <https://bit.ly/3LxoPUx>

James Grimmelmann (james.grimmelmann@cornell.edu) is the Tessler Family Professor of Digital and Information Law at Cornell Tech and in the Law School at Cornell University, New York, NY, USA.