

Blocking the Internet Archive Will Not Stop AI

Brewster Kahle

In 1996, during the early years of the web, I could see that those of us who were participating online were creating an extraordinary new record of human activity—and allowing it to disappear almost as quickly as we made it.

The web was becoming the place where government officials made promises, journalists published information in the public interest, and ordinary people shared ideas and documented their lives. Yet it was fragile—equal parts consequential and ephemeral.

Many people assume that once something is published online, it will remain forever retrievable. The saying that “the internet never forgets” has produced a dangerously false sense of permanence. In reality, the internet forgets continuously. Websites are constantly revised, deleted, and abandoned altogether.

This is why, 30 years ago, I founded the Internet Archive. I wanted to build a mechanism for permanence in the internet age. Soon after that, I helped build the Internet Archive’s most-used resource, the Wayback Machine, a system that would preserve webpages as they appeared at any one moment in time, allowing future readers to return to those moments. The idea was simple but urgent: If the internet was going to become humanity’s principal publishing platform, then it needed what every great publishing system that came before it had needed. It needed a library.

People all over the world use the Wayback Machine all the time. What began as an effort to preserve the web has become an essential tool for journalists, researchers, historians, and the broader public.

Yet the future of this tool has become extremely vulnerable. *The New York Times* is now blocking the Internet Archive’s ability to preserve its articles. *The Guardian* has restricted our access to its published articles and removed articles from some of the Wayback Machine’s interfaces. *USA Today* and a growing number of other news organizations have also placed barriers in the path of web archivists (even, yes, *The Atlantic*). As of May 2026, Harvard’s Nieman Journalism Lab had identified more than 340 local-news sites that were limiting the Internet Archive’s ability to preserve their journalism.

The publishers taking these steps have all pointed to one reason: the unauthorized use of their work by AI companies to train large language models. They complain that our library serves as a back door around paywalls, licensing negotiations, and technical restrictions. The *Times* has said that it is blocking us because the Wayback Machine allegedly provides “unfettered access” to its work. But this fear-based approach does not take into account that the Internet Archive is taking steps to prevent AI companies from scraping the materials we preserve.

I understand why publishers are angry. Journalism is expensive. Reporting requires skilled people, time, courage, legal support, and institutions willing to withstand pressure. Journalists have watched advertising revenue disappear and subscription growth slow while technology companies build enormously profitable products using material created by others. Publishers have every right to challenge commercial businesses that they believe are inappropriately exploiting their work.

But blocking a nonprofit library is the wrong response to this problem. Impeding the Wayback Machine’s preservation work will not prevent powerful AI companies from gathering training data. It will certainly not restore the advertising-based business model that supported newspapers in the 20th century, a business model that has long since collapsed. What it will do is prevent historians, researchers, courts, ordinary readers—and, most ironic, journalists themselves—from consulting an accurate record of what once

appeared on the web. It is difficult to quantify the scale of what we stand to lose. A continuous record of web activity is an enormously powerful accountability tool, especially at a time when seeking the truth and reporting it has become ever more complicated and the informational environment is growing only more complex.

Joe Mullin, who is a senior policy analyst at the Electronic Frontier Foundation, has offered a fitting analogy: Imagine a newspaper announcing that physical libraries will no longer be permitted to preserve copies of its past issues. That is effectively what is happening now. For the past three decades, the Wayback Machine has served as a library for the public web. The destruction of that library is creating holes in the historical record.

A crucial fact in this debate is one that publishers themselves know better than anyone: Journalists rely on the Wayback Machine. They use it to recover deleted tweets, verify past public statements by powerful individuals who wish to avoid scrutiny, examine earlier versions of government documents, and establish what a person or institution said about a subject before some controversy or scandal prompted them to change course. Journalists use the Wayback Machine to see defunct websites and the archives of closed newsrooms. They use it when the primary evidence for a story has vanished from the web.

A news organization that blocks the Wayback Machine may believe it is acting in the best interest of its journalists. But instead it is dismantling one of journalism's most important investigative instruments.

My argument is simple: The public record needs an independent steward. A society that cannot inspect its past is easier to deceive. In 2017, Katie Hafner and Griffin Palmer, writing for *The New York Times*, relied on the Wayback Machine to report that private-equity investment was reshaping dermatology practices—expanding the use of physician assistants, increasing the volume of biopsies and procedures, enacting ever more aggressive screening and treatment protocols, and exposing elderly patients to treatments whose benefits were doubtful.

The Wayback Machine gave Hafner and Palmer an empirical record that proved crucial for their story: evidence of change over time. The current website of Advanced Dermatology and Cosmetic Surgery (one of the companies they were looking into) showed a sprawling operation with more than 180 locations and 124 physician assistants. But earlier versions of the site showed that the number of physician assistants listed by the company had increased by 400 percent since 2008, amid a debate over whether physician assistants are qualified to properly make diagnoses. When I spoke with Hafner about the story, she told me she could not have made her point as powerfully as she did without the Wayback Machine. The records she retrieved from our archive enabled her to demonstrate the scale, speed, and direction of a corporate transformation, including the rapid growth of a business model that placed many more lesser-trained practitioners in clinics serving vulnerable patients.

Hafner and Palmer are among thousands of reporters whose use of the Wayback Machine has contributed to extraordinary journalism. A politics reporter may find that an agency quietly removed a document, or even a single sentence, in the face of criticism. A city-hall reporter may need an earlier version of a municipal page containing budget figures, meeting records, and public-health guidance. A technology reporter may need to compare successive versions of a platform's terms of service. An investigative journalist examining allegations against a public figure may need to scour a decade-old webpage that has since been deleted.

Government agencies routinely scrub important information from their sites, and the Wayback Machine is a faster and more practical method of discovery than a Freedom of Information Act request. These are not hypothetical benefits. Over and over and over again, the Wayback Machine has confirmed facts that would be irretrievable without an independent archive.

Our own monitoring finds that more than 100 news articles in a typical month cite or rely on material preserved by the Wayback Machine. We can reasonably assume that there are tens of thousands of additional visits to the Wayback Machine every month, when an archived page might help illuminate the truth.

Mark Graham, who directs the Wayback Machine, recently described having visited *The New York Times* to meet with data journalists and reporters—and being approached by a senior researcher who thanked him profusely. The researcher said Times staffers used the Wayback Machine constantly and had even found material there that they could not find anywhere else. Consider what that means. Even one of the best-resourced news organizations in the world does not possess a complete, easily usable record of the material it needs.

Plenty of journalists—including [at The Atlantic](#)—have championed and defended the importance of the Wayback Machine. MS NOW’s Rachel Maddow has called the Internet Archive a national treasure and said she uses it daily. She and hundreds of other journalists have signed a statement calling for web preservation.

These journalists understand that their work is cumulative. A story links to documents, earlier articles, public statements, data sets, and numerous other verifiable facts in the public record. Future stories return to (and build upon) that evidence. When underlying links die, one of the chains of verification breaks along with it.

In 2024, the Pew Research Center announced the findings of a study that tracked 1 million webpages over a 10-year period. At the end of the study, one-quarter of those pages were no longer accessible. Among pages that existed in 2013, the first year of the study, the loss rate was 38 percent. Even among pages collected as recently as 2021, roughly one in five had disappeared within two years.

This is not the timescale of ancient history. But it is often the timescale of someone’s political career, a long-running lawsuit, or a company’s rise and fall.

Elements of the web disappear for all kinds of reasons, nefarious and mundane. Perhaps a page is deleted during a redesign. Maybe a publication changes ownership, or a content-management system is replaced.

After the Pew study came out, the Internet Archive took a closer look at its data set to determine how much of the lost material had been preserved by the Wayback Machine. The analysis found that 16 percent of the vanished URLs are accessible via the Wayback Machine.

Consider the *Rocky Mountain News*. Colorado’s oldest newspaper published for nearly 150 years and earned a reputation for forceful accountability journalism that exposed corruption, misspent public money, and institutions that were failing the people they were meant to serve. In its final decade, the paper won four—I repeat, four —Pulitzer Prizes: for photographic coverage of the Columbine school shooting, for its images of Colorado’s devastating wildfires, and for the writing and photography of “Final Salute,” a deeply reported account of Marines killed in Iraq. In 2015, *The Atlantic*’s Adrienne LaFrance [reported](#) the story of how one of the paper’s Pulitzer finalists successfully resuscitated an investigative series of his that had vanished from the web when the paper did. Her story concluded with a link to the series that he’d saved. Even *that* link is now dead. (Though you can still find the series [here](#).)

No institution, however prestigious, is immortal. The Wayback Machine holds the work of publications that no longer exist, such as the *Rocky Mountain News*. It preserves the archive of MTV News and BET.com, including decades of cultural reporting removed from the web without warning. It preserves a robust record of early-2000s web culture from Gawker Media. It provides access to work produced by Apple Daily and Stand News, Hong Kong-based news outlets that were forced to shut down in 2021 as part of a government crackdown, as well as independent Russian-language journalism, some of it produced in exile. It holds the only copy of *USA Today* from September 11, 2001, because the paper’s website was overwhelmed and its online editors continuously updated stories by hand without preserving the changes made along the way.

The Internet Archive is not an AI company. A commercial AI laboratory, a search engine, a research project, and a nonprofit library may all use automated systems to retrieve publicly available webpages. From a server’s perspective, each may appear to be a bot. But not all bots are bad bots. They do not have

the same mission, business model, or conduct.

Publishers' concerns about abusive scraping are merited. I know this acutely because the Internet Archive has itself been [subjected to extreme machine traffic](#). In one incident, an AI startup [sent tens of thousands of requests](#) per second and overloaded our systems. We blocked the offending hosts, contacted the company, and worked to stop the behavior.

When a company attempts to extract material at scale, it burdens our nonprofit infrastructure and interferes with access by ordinary users. We understand why publishers object to how AI companies have used their work. But recognizing a problem does not justify attacking the wrong institution and disregarding the public interest. In short, the Wayback Machine has become collateral damage in a fight between publishers and AI companies. Blocking the Internet Archive will not stop AI. Commercial AI companies have crawled the live web, licensed large data sets, purchased access from platforms, obtained third-party collections, and built their own retrieval systems. Publishers can sue those companies, negotiate with them, or build technical controls specifically directed at them. We have already built such controls to protect the material on our site—including the material we preserve from news organizations.

Web archiving belongs to a well-established legal tradition. Courts have repeatedly recognized that copying material for indexing and search can be a transformative fair use. A searchable system cannot function without making technical copies of the material it indexes. In the now-famous Google Books litigation, in which authors sued Google over the digitization of millions of books, courts held that scanning books to create a searchable research tool served a different purpose from selling or reading the original books, and was fair use under U.S. copyright law.

Defending preservation does not require ignoring publishers' legitimate concerns. The Internet Archive has adapted repeatedly as the web has changed. We use rate limits, filtering, monitoring, and network-security systems to respond to abusive access. We have restricted bulk functionality in particular contexts and optimized important parts of the Wayback Machine for human use.

Those efforts can and should continue. Publishers and archives can distinguish ordinary human consultation from sustained bulk extraction. We can develop better methods for identifying abusive traffic, create responsible-access arrangements, improve communication during high-volume incidents, and discuss access to application programming interfaces separately from the preservation of publicly available pages. Most important, we can preserve records while applying reasonable protections against mass downloading.

The press cannot have it both ways. Journalists describe their work as the first draft of history—but a draft of history must survive if it is to be revised, improved upon, and understood.