Can You Prevent AI From Scraping Your Website Data? District Court Says Answer Lies in Copyright Law

Article By:

Dan Jasnow

Danielle W. Bulger

As the prevalence of artificial intelligence (AI) continues to rise, complex questions regarding the regulation of AI data scraping remain relevant to both website owners and web data collection companies. Though many websites have attempted to prohibit AI data scraping of their content through their terms of use, a federal court clarified this year that the extent to which public data may be scraped from social media platforms should be governed by federal copyright law.

Data scraping is the process by which data is extracted and copied from websites, including social media platforms, usually for the purpose of training AI large language models. Because scraped data can include user-generated content and personal information, many companies explicitly prohibit data scraping through their terms of use. One such company is X Corp., owner and operator of the social media platform X (formerly known as Twitter), which recently attempted to enforce its anti-scraping policies against data-scraping company Bright Data Ltd. in federal court.

That suit made headlines when the US District Court for the Northern District of California used federal copyright law to side with Bright Data, dismissing X Corp.'s breach-of-contract claims against Bright Data. *See X Corp. v. Bright Data Ltd.*, No. C 23-03698 WHA, 2024 WL 2113859, at *14 (N.D. Cal. May 9, 2024).

X Corp. sought to establish liability for Bright Data accessing X Corp.'s systems to scrape and sell data from the X website. X Corp. claimed that Bright Data was bound by "browsewrap" and "clickwrap" agreements that Bright Data agreed to when it created its X account (@bright_data). Both agreements explicitly prohibited scraping of X Corp.'s data and, therefore, X Corp. argued that Bright Data was liable for breach of contract, tortious interference with a contract, and unjust enrichment. Surprisingly, the court disagreed, noting that X users — not X Corp. — own and retain rights in content posted on the X website, thereby invoking the Copyright Act, preempting X Corp.'s claims, and justifying dismissal of the complaint for X Corp.'s failure to state a claim.

In its decision, the court reiterated terms outlined in X Corp.'s Terms of Service, which expressly state that X users, rather than X Corp., retain copyright ownership in all information, text, links, photos, and other materials and content they submit, post, or display on the site. Under the Terms, X users grant to X Corp. a non-exclusive, royalty-free license to use, copy, display, and distribute the content, among other tasks. Given X Corp.'s limited role as a non-exclusive licensee, the court admonished X Corp. for improperly attempting to expand its rights, noting such an expansion would exclude others from using and distributing X users' content. The court further criticized X Corp. for attempting to enact its own private copyright system that conflicts with US Congress's copyright system, noting that the extent to which public data may be freely copied from social media platforms should generally be governed by the Copyright Act.

The court's decision emphasizes the purpose and intended effects of the Copyright Act, including the broad rights of copyright owners, regardless of the content's medium. As explained by the court, the Copyright Act empowers X users as copyright owners — not non-exclusive licensee X Corp. — to exclude others from exploiting their exclusive rights. This is despite the fact that X users choose to license their copyrighted content with the purpose of making it globally accessible via the X platform. The decision also highlights the importance of the statutory privilege of fair use in a digital context and the efforts courts are taking to ensure content Congress intended to be free from copyright protection — including website likes, usernames, and short comments — remains free from restraint.

Although the court dismissed this case, Judge William Alsup clarified that not every state law interest would be automatically preempted whenever such interests burden copyright protections. Judge Alsup advised that, for example, analogous state law claims designed to protect social media users' privacy should not be preempted by the Copyright Act because privacy protection is not a function of copyright law.

Nevertheless, for companies in the business of data scraping, this decision likely comes as good news given that the court refused to hold a data-scraping company liable for breach of a website's Terms of Service that expressly prohibited scraping. In light of the court's decision, data-scraping companies should conduct copyright analyses of the data they wish to scrape to determine whether they can be held liable. These analyses will turn on the originality of the content and whether the fair use defense applies.

As for website owners, social media companies, and internet service providers, this case serves as a reminder of the Copyright Act's limitations imposed on them by virtue of being non-exclusive licensees. The case also illustrates that terms of use will not be sufficient to protect against data scraping as long as the underlying data copied is either subject to fair use or not protectable under the Copyright Act.

Either way, copyright law and copyright analysis of content remain central to all parties impacted by the practice of data scraping. Ultimately, however, how infringement and fair use is analyzed in the context of certain data scraped, particularly by AI, remains a question before many courts nationwide.

Source URL:https://natlawreview.com/article/can-you-prevent-ai-scraping-your-website-data-district-court-says-answer-lies