



Assessing the Risk of Web Scraping

Tech Series: what's web scraping?

November 2, 2023

By [Andre Assumpcao](#)

Web scraping occurs in many public-facing court records and docket sheets. It has negative ramifications for those whose banks or employers [incorrectly use those records](#) in background checks. It can also slow down the courts' websites for those who need to access them. It can also slow down the courts' websites for those who need to access them.

Web scraping is extracting data from websites using automated tools or programs to access web pages, read source code or other structured data, and extract the desired information.

Some common uses for web scraping are market research, price comparison, academic research, and court data collection. Surety businesses, debt collectors, and background check companies are some examples of parties interested in obtaining court data. Web scraping can be done over public or proprietary data available on the web; thus, it is important to know that web scrapers will not necessarily consider [ethical](#) and legal implications of collecting and using web data.

Web scraping causes concerns for courts. It might slow down a court's website, expose personal identifiers or other information from internal systems (i.e., scrapers that access content behind a firewall or login and expose usernames and passwords), infringe on intellectual property (i.e., collection of proprietary content), and misrepresent data (i.e., data extracted without context or altered and manipulated).

Despite lawsuits, web scraping is [legal](#) so preventing it is difficult. There are ways to mitigate the risk of web scraping, however. Options include posting clear terms of use for your website; posting actionable legal guidelines for those who scrape and misuse your data; implementing rate-limiting technology to reduce the ability of web scrapers to quickly obtain data; implementing database URL masking to prevent scrapers from automatically visiting pages that encode information in their web address; adopting CAPTCHA technology for accessing data; providing data in bulk form via official request (with or without a fee); and finally offering data through an application programming interface (API), thus making web scraping moot.

There are situations in which web scraping presents minimal risk and should be considered as a data collection tool. These factors depend on whether the data are proprietary or public; the data contain personally identifiable data; the website has heavy or light data traffic; the purpose of the data (i.e., research, public good); and whether the source code of the website prohibits or allows data scraping through the robots.txt hidden file guidelines. Learn more at [NCSC Data Dives](#).

Does your court have policies concerning web scraping? Email us at Knowledge@ncsc.org or call 800-616-6164 and let us know. Follow the National Center for State Courts on [Facebook](#), [X](#), [LinkedIn](#), and [Vimeo](#). For more Trending Topic posts, visit ncsc.org/trendingtopics or subscribe to the [LinkedIn newsletter](#).