

5 Things to Know About Data Scraping – A Blog Post by David Goldenberg

How you build your data, and where you get it, can create legal issues for your company. How can you prevent others from taking your data? A summary of the law and a few tips below.

Data scraping is a concept that has been extremely controversial in the last few years. Briefly, data scraping is the extraction of data from a website or other place on the internet, typically via an automated process, such as a script or scraping bot. The bot will typically look for structured data on the third-party site and extract and load that data based on the parameters set in the program.

Data scraping is extremely common and may account for over a quarter of all website hits. These bots can be considered good or bad, depending on how the extracted information is used. However, just because you can technically gather this data, does not mean it is legal to do so. As technology evolves, are the laws surrounding this type of activity capable of keeping up?

1. Data Scraping is Useful

Data scraping of websites is actually an enormous part of how the internet works. These programs are used to gather data, such as search engines evaluating web content or collecting information on prices for comparison purposes. Many of these uses are not controversial and have been occurring for years.

2. Good Bots vs. Bad Bots

However, at the same time that all this permitted activity has been occurring, there has been a parallel battle between content creating / hosting companies and those that want to scrape this content. Often it is how the scraper uses the information which will determine whether the host site will want to prevent the activity. For instance, most commercial websites actively want Google and other search engines to take the information from their page to present their website in search results. However, other bots may take the information to use on a competitive site, or worse, for account theft or ad fraud. It has been estimated that bad bots make up over 20% of all web traffic.

3. Data Scraping Protection Isn't Common

As mentioned above, data scraping isn't always a bad thing, and most websites allow what they perceive to be beneficial scraping to occur. However, most websites aren't equipped with tools to stop unwanted data scraping. Stopping data scraping isn't the same as having virus protection software, so it requires more thought and planning to prevent. Surprisingly, there is a lack of technological services for this need. There are companies that can prevent bad bots, but there aren't many fully automated programs that can handle the onslaught of page hits. In addition, many companies do not explicitly consider scraping when crafting their terms of service.

4. The Laws May Lag Behind the Technology

A recent case regarding web scraping has renewed interest in scraping tools and what can be done to prevent it. In general, a website's terms of service and copyright law are the primary legal tools to stop scraping. In the summer of 2017, hiQ Labs, a startup based out of San Francisco, sued LinkedIn, the well-known online resume and profile site. hiQ was scraping available LinkedIn profiles to offer potential clients insight into certain applicants (how often they switched jobs etc.). On August 14th, 2017, Judge Edward Chen ruled that LinkedIn had to remove the barriers preventing their site from being scraped by hiQ. Interestingly, it appears that LinkedIn's terms of service did not clearly prohibit the activity. In 2019, the 9th Circuit affirmed the district court's preliminary injunction, preventing LinkedIn from denying hiQ Labs from accessing LinkedIn's publicly available LinkedIn member profiles.

5. Data Scraping Isn't Going Away

While laws are currently unclear about the exact parameters on which type of data scraping is legally allowed, a strong and clear set of terms preventing unwanted activity will provide strong rights against unwanted scraping.