# Web Scraping

Web scraping is the process of extracting data from websites. It’s been around since the beginning of the web. It can be done manually. There’s nothing stopping anyone from going in and copying and pasting every web page on a website and placing it into an Excel spreadsheet or a database. There’s been a command line tool called wget issued with most Linux distributions for over twenty years. It has a recursive option that downloads all the subdirectories of the website — everything on the website — and replicates the original site. There are sophisticated tools for web scraping now. Some are freeware and some are commercial. Artificial intelligence (AI) is part of the toolset as well now.

Web scrapers are usually looking for something specific on a website, and these days web scraping is usually executed by bots with the entire site or specific pages, delimited by data or data markers, selected by a web crawler and downloaded. There’s often a post-processing stage in which the very specific data sought from the pages is extracted and placed into a meaningful structure in a database.

The most distinct commercial use for web scraping is price comparison and price fluctuations among competitors in an industry. Real estate listings and airline fares are examples of data commonly targeted for scraping. It’s also used for getting client lists or lists of people who might be interested in a service or a product as demonstrated by their signing up on a website, all of which can be extracted and harvested through web scraping.

Like all things on the web, the process of web scraping has evolved. Data feeds between web pages and website databases can be targeted now. The data coming out of the database is often transported by JSON and can be intercepted and extracted. This is where things start to enter into a legal ambiguity. When you put a website up you have no control over who can download the pages. The dynamic process by which those pages are populated, based upon interactions with users of the website, is proprietary, and, as such, violates privacy and data security more transparently.

The use of web APIs to facilitate data transfer has added to and compounded the problem of web scraping, which can be further refined with scripts using the regular expression syntax as well as things like the Unix grep syntax. These syntaxes allow web scrapers to pinpoint their scraping and extraction. Dynamically generated pages on websites are almost always populated by a database. The data categories are often encoded by a script or template. Sophisticated web scrapers, though it might be better to call them data miners, can write programs that find the templates in an information source, extract its content, and translate it into a readable form. In the scraping lexicon, this is called a *wrapper*. Even more sophisticated forms of web scraping and data mining use data query languages like HTQL to rapidly go through web pages, get data, and then present that data in a usable format. Harvesting platforms have been developed by several companies, capable of creating, deploying, and monitoring their bots. These platforms go after an entire vertical of information on the web.

There is no global consensus on the legality of web scraping. Different national and state or provincial jurisdictions have different approaches to it. In the United States, a victim of web scraping and commercial loss as a result of that web scraping could file a claim using copyright infringement, a claim under the Computer Fraud and Abuse Act, or simply crimes against property. The success of this approach to defense against web scraping depends on certain conditions. If the victim uses the copyright approach to the legal remedy it has to be remembered that case law dictates that the duplication of facts is not illegal. So then the victim of the web scraping is responsible for demonstrating to a court how proprietary data or data beneficial to their revenue stream was taken from them. Some courts in the US have accommodated claims against property. This is almost like trespassing, but it takes place in the digital world, and it relies on the fact that the computer and the computer system are the personal property of the website host. These issues become particularly testy and litigious in markets that rely on an intimate understanding of price scale on a day-by-day, week-by-week, and month-by-month basis. Airline fares are a good example, resort and vacation offerings as well, and for the same reasons.

So how do you stop it? Most of the web scraping now is done by bots. Some of the bots declare themselves, and some do not. The most immediate defense is to allow verification of human use at certain levels of engagement on the website. This can be facilitated by sophisticated defense programs like hCaptcha, which combines machine learning (ML) with a simple capture functionality employing image recognition in a matrix. The ML function is mandatory now. Many bots now have their ML that can circumvent simple image recognition designed to verify human interaction. As one side goes up, whether it be the side of security or the side of fraud, the other must reciprocate, and the new standard in human verification must include ML to stop the comparably evolved bots. Web scraping can cause damage to the bottom line of any business.