Web Scraping: IP Considerations & Best Practices
October 21, 2016
# Table of Contents

- Why This Topic ........................................ 3
- Methodology ............................................ 3
- Definition and Use Cases ............................ 4
- Legal Considerations .................................. 5
- Selected Legal Cases ................................... 6
- Essential Actions and Best Practices ............. 7
Why This Topic

Web scraping is an important tool for information managers, researchers, scholars, journalists, and businesses of all sizes to gather large quantities of rich data for analytics purposes. The sources of scraped material are myriad, ranging from company websites to social media to government data sets, as are the reasons for conducting web scraping.

Ultimately, web scraping activity often sets the interests of a website owner against those who want to harvest the data. Owners usually want to control, profit from, or leverage the data they have on their website. Those that want to collect data may want to use it for other purposes such as competitive advantage, analysis, profit, or reasons completely unconnected to the original website. As utilization of big data analytics to find patterns and trends hidden in unstructured or seemingly unrelated data sets continues to grow, the degree and volume of web scraping will also continue to increase. Where many of the first court cases involving web scraping centered around simple copying, republishing, or linking to scraped data for competitive purposes, the use of data analytics tools has broadened the playing field of possible uses of scraped material, and it is likely to make interpretations of liability more nuanced and driven by the circumstance of each specific case.

In this briefing, we look at the definition of web scraping and some use cases in current practice, followed by a review of the legal considerations and theories of liability through which website owners have tried to sue or prevent web scraping activity. We also present some sample cases illustrating those theories. Finally, we look at some best practices for companies that use or are considering the use of web scraping as part of their business intelligence practice.

Methodology

Primary research for this report is comprised of a series of 10 interviews with information managers, business intelligence researchers, and legal practitioners. We supplemented those interviews with secondary research from published reports, industry blogs, mainstream press, and blogs. Outsell’s ongoing dialogue with the market added depth to the primary and secondary research.
Definition and Use Cases

If web content can be viewed on a page, it can be scraped.

- Web scraping, sometimes called web harvesting or web data extraction, is a software technique used to copy or collect selected information from websites. Automatic crawling of websites is widespread, with “crawlers” (or “robots” or “spiders”) extracting varying amounts of data and information.

- Companies use web scraping for a variety of reasons: harvesting publicly available data like financial, weather, news, company information, sports scores, etc. Others use it to watch trends, gather competitive intelligence, product information, monitor changes in websites, or strategic insight. Web search engines use a similar technique to index websites, capturing links and short snippets of information, in order to catalogue the data and provide accurate search results.

Sample Use Cases

- **Real Estate:** Scraping housing sale or rental data from online real estate listings in order to track pricing trends over time, average prices, or uncover indicators useful to real estate agents, builders, investors, etc.

- **Legislative or Regulatory:** Scraping government or regulatory agency websites to discover, understand, or monitor changes or developments with legislation or regulations.

- **Health Care:** Scraping healthcare forums and blogs to gather conversations or comments about particular drugs, medical devices, or treatments in order to monitor brand reputation, adverse effects, or other clinical information.

- **Public libraries:** Archiving and preserving local or commercial history.

- **University libraries:** Archiving course syllabi and campus activities.

- **Internet Archives:** Compiling free books, movies, music, software, etc., or simply archiving website to compile a history of the web.

Web scraping is also employed in any number of illicit or hacking applications. Corporate espionage, theft of intellectual property, wholesale copying of websites, or impacting website performance, are all examples of employing scraping with an intent to cause damage or loss.
Legal Considerations

The legal landscape around web scraping is still in its early stages and continues to evolve. Under certain circumstances, outlined below, the courts in the US have found that the activity is legal and is something that website operators should expect.

By some estimates there have been less than 100 cases involving web crawling in the US. One of the reasons is that the likelihood of a case being successful for the plaintiff is difficult, as is the ability to obtain compensation in the event of a favorable judgment. Therefore, website operators are reluctant to spend on legal resources, even if there is clearly a web crawler violating their site.

There are several theories of infringement or liability, by which owners of websites have brought legal action against web scrapers. However, as already mentioned, the case law is still very much in flux. The most prominent cases involved implementing one or more of the following:

**Breach of terms and conditions**

Many websites post terms and conditions (T&Cs) prominently on their pages that address the issue of access to their website via scrapers. This is intended to create breach of contract liability by establishing a contract between the website owner and the scraper. Posting T&Cs (also known as a “browse-wrap”) may not be enough to show that a scraper has breached terms, since there is no active acceptance on the part of the scraper. What appears to be more enforceable is the use of a “click-wrap” in which the web scraper has to actively click to accept the T&Cs, meaning there is proof of acceptance of the T&Cs.

**Copyright or trademark infringement**

US courts generally consider a number of factors when it comes to deciding whether scraping and using website content is fair use. Among the factors are:

- How the scraper uses the content, and whether or not it is for personal or commercial use.
- The type of material scraped.
- The volume of the material scraped.
- The market value or potential value of the material.
- Whether or not the material is simply reproduced, or if it is transformed or enhanced in some way.

**Computer Fraud and Abuse Act (CFAA)**

There are several federal and state laws against hacking or accessing another’s computer. The CFAA states that whoever “intentionally accesses a computer without authorization ... and as a result of such conduct recklessly causes damage” has acted in violation of the act, especially if the violated website can prove loss or damages.

**Trespass to chattels**

This is a term referring to a civil wrong which means one entity has interfered with another’s personal property which causes loss of value or damages.

**Robots exclusion protocol**

This is an industry standard program that allows a website to embed “robots.txt” files within their websites that communicate instructions to web crawlers to indicate which crawlers can access the site, and which pages they can access. While a common protocol, it has limited legal value, mainly as supporting evidence in cases involving breach of terms and conditions.
Selected Legal Cases

While theories of liability are still developing, the pattern suggests that the courts are leaning towards protecting proprietary content, especially when there is some indication of damages suffered by the site owners. Circumstances vary in each case and depend on how scraper accessed content, what the scraper harvested, and the terms of use presented by the website owner. The following are a sample of issues and resolutions illustrating the range of liabilities.

**Associated Press vs. Meltwater**

This high-profile case was decided for Associated Press (AP). The media-monitoring firm Meltwater had been crawling AP websites and had extracted and republished significant amounts of text from AP news articles “without adding any commentary or insight.” Meltwater’s stance was that it was operating under fair use provisions of copyright laws. Although the court found in favor of AP, the firms settled the case and entered into a partnership arrangement to develop new products.

**Ticketmaster vs. Riedel Marketing Group**

Riedel Marketing Group (RMG) used web robots to saturate the Ticketmaster site so that it could harvest large quantities of desirable tickets for resale. Ticketmaster argued that RMG had agreed to the terms and conditions of the site but ignored them. Ticketmaster obtained a preliminary injunction against RMG, and the court held that RMG had infringed on Ticketmaster’s copyrighted material.

**eBay vs. Bidder’s Edge**

Bidder’s Edge used a web scraper to obtain details of active eBay auctions and reposted them on its own site. In addition, it crawled eBay’s site more than 100,000 times per day, which hindered eBay’s site performance. Bidder’s Edge ignored the robots.txt files as well as a cease and desist letter from eBay, and eBay brought a trespass to chattels claim against Bidder’s Edge. eBay successfully obtained a preliminary injunction against Bidder’s Edge, since the court decided that Bidder’s Edge spiders interfered with the capacity of eBay’s servers.

**QVC vs. Resultly**

QVC’s terms and conditions did not prohibit web scraping, but Resultly was so aggressive with its scraping that QVC’s servers became overburdened, and customers were unable to complete transactions. QVC alleged a loss of $2 million, and sued Resultly under the Computer Fraud and Abuse Act (CFAA), seeking a preliminary injunction. The court ruled that the plaintiff had to prove that the defendant “both knowingly and intentionally intended to cause damage to the plaintiff’s computer.” In this case the court found that Resultly did not intend to cause damage, so QVC was unable to show that Resultly violated the CFAA.

**Craigslist vs. Naturemarket**

Due to the extensive amount and ever-changing listings on Craigslist, the site has been the target of numerous scrapers over the years, and it has successfully sued or stopped numerous organizations from scraping data. In one case, Naturemarket sold software that made it easy for customers to automatically post listings to Craigslist and to post listings on behalf of customers. Naturemarket also scraped email addresses from Craigslist’s site. Craigslist sued, using copyright infringement, the Computer Fraud and Abuse Act, and breach of terms of use. Naturemarket failed to contest the suit, and the court awarded Craigslist a judgment of $1.3 million.
Essential Actions and Best Practices

Web scraping can be a useful and constructive tool for gathering unstructured information to use for insights on multiple levels. While the legal landscape remains unsettled, and uncertainties abound in terms of what is permissible with web scraping, in Outsell’s opinion there are legal, ethical, and social boundaries that are important to maintain. Companies that conduct scraping activities must carefully consider the consequences of their actions. The following items are best practices to limit risk or liability when conducting web scraping:

- **Comply with terms and conditions and robots.txt**: If there is an advisory in the form of a robots.txt file, or terms and conditions specified which permits, limits, or prohibits web scraping, it’s best to stay in compliance and not breach the implied contract.

- **Don’t overburden the website**: Make sure that excessive queries don’t interfere with a website’s normal processes, slow a site’s performance, or cause unintended harm to the website operator.

- **Copy as little as possible**: Avoid making complete copies of web pages or design elements from the site. Focus on factual information such as pricing data, dates, locations, etc. When it comes to non-factual information, store it only as long as required for analysis. One company that we interviewed has a policy of keeping non-factual, competitive information for only 30 days, during which they analyze it, and then discard it.

- **Don’t plagiarize**: Don’t take information from a site and redistribute it or use it verbatim. Transform it, analyze it, or use it in different ways than the original site.

- **Avoid crawling or storing sensitive data**: Certain types of non-public information or personally identifiable information, such as names, birth dates, or email addresses, are off-limits.

- **Don’t compete**: Avoid using scraped data to offer a competing product or service, which can potentially cause damage to the scraped website.

- **Be transparent**: Scraping technology is such that it’s possible to put your company’s contact details in the scraper’s header to identify who is doing the scraping. Using technical trickery to mask IP addresses, bypass CAPTCHAs, or otherwise conceal or filter activities is off limits. Don’t misrepresent one’s identity or purpose, or use deception in order to gain access.

- **Use an API if it’s there**: Many websites offer the ability to download data via an API, either free or for a fee. If that is the case, it’s preferable to obtain the data that way, rather than by scraping.

- **Obey cease and desist requests**: If a website operator wants the scraping to stop, acknowledge and stop. In the long run, it’s better to adhere to the request than to risk liability.

- **Keep up-to-date**: The legal landscape is still in its formative stages and there are no clear-cut patterns in the decisions. In addition, intellectual property laws are not uniform among various jurisdictions or countries. It is important to keep abreast of the developing case law centered on web scraping to mitigate risk.

- **Develop a written company policy**: Web scraping is a normal and expected business activity. It’s not possible to eliminate risk altogether, but establishing and adhering to clear guidelines for web scraping activities is critical from both a legal and ethical standpoint.
ABOUT THE AUTHOR

Jim Hydock
Affiliate Analyst
+1 650-342-6060
jhydock@outsellinc.com

ABOUT OUTSELL

The rapid convergence of information, media, technology, and data is reshaping businesses every day. Enter Outsell, Inc., the only research and advisory firm focusing on these four sectors. As the trusted advisor to executives, our analysts turn complexity into clarity, and provide the facts and insights necessary to make the right decisions. Our proven blend of big data, research, proprietary intelligence, and exclusive leadership communities produces tangible results and a strong ROI. We promise to deliver “wow” and ensure our clients stay more focused, save time, and grow revenue in a fast-changing digital world.

www.outsellinc.com
contact_us@outsellinc.com

Burlingame, CA USA
+1 650-342-6060

London, United Kingdom
+44 (0)20 8090 6590

Outsell, Inc. is the sole and exclusive owner of all copyrights and content in this report. As a user of this report, you acknowledge that you are a licensee of Outsell’s copyrights and that Outsell, Inc. retains title to all Outsell copyrights in the report. You may use this report, only within your own work group in your company. For broader distribution right options, please email us at info@outsellinc.com. The information, analysis, and opinions (the “Content”) contained herein are based on the qualitative and quantitative research methods of Outsell, Inc. and its staff’s extensive professional expertise in the industry. Outsell has used its best efforts and judgment in the compilation and presentation of the Content and to ensure to the best of its ability that the Content is accurate as of the date published. However, the industry information covered by this report is subject to rapid change. Outsell makes no representations or warranties, express or implied, concerning or relating to the accuracy of the Content in this report and Outsell assumes no liability related to claims concerning the Content of this report.

Advancing the Business of Information
© 2016 Outsell, Inc.