Intellectual Property Today™

Spiders, Crawlers and Bots, Oh My: The Basics of Website Scraping

BY PERRY J. VISCOUNTY, JENNIFER L. BARRY AND JAMES FIELD OF LATHAM & WATKINS

In an increasingly digital age, a greater and greater percentage of business is conducted online. Often, a company's website is the primary conduit through which a consumer interacts with the company. While this accessibility to consumers has undoubtedly been beneficial to many businesses, the amount of information available to the public via such websites has grown commensurately. This growth has in turn led many businesses to face an unanticipated challenge: the increasing incidence and growing complexity of web scraping.

Web scraping (also called web harvesting or web data extraction) is a pervasive and increasingly sophisticated practice used to extract information or data from third party websites, usually with the intent to use that data for commercial purposes. In the process of web scraping, software programs (sometimes called "bots," "crawlers," or "spider" programs) are often used to simulate human exploration of the web by automating the process a human would follow in exploring the content of a site. Among the myriad of applications (both legal and otherwise) of web scraping technology, some common commercial uses have developed. Examples include the creation of websites that allows users to search for travel rates from multiple third-party sites in real time; websites containing information on products for sale collected from numerous other websites; and mining text headlines from news websites to create an aggregated news feed.

Depending on the scraping method employed and the use of the scraped information, this mining of data may trigger liability to the party conducting the scraping. Scraping can trigger liability in a number of different ways, including violating the scraped website's terms of use, infringing trademarks, or otherwise damaging the economic interests of the website owner. Consequently, website owners need to vigilantly monitor and secure the data on their websites, and companies contemplating scraping activities should carefully consider the legality of that activity.

There are several different theories of liability that can be applied to a scraper's use of another's website or software to gather and extract information, including: breach of contract; trespass to chattels; unfair competition; and violations of federal computer fraud statutes. Some of these theories have been successfully utilized since web scraping first began to occur; others are relatively new additions to the arsenal in combating illicit web scraping. This article explores some of the popular and cutting-edge theories of liability for web scraping and gives practical advice for protecting the content of your personal or professional websites.

TRESPASS TO CHATTEL

One of the more recent theories of liability for web scraping activities (and perhaps the most inventive) is trespass to chattel. Courts have begun to apply this common law cause of action to situations involving electronic interference with, or unauthorized access to, the computer systems of a third party. To succeed on a claim for trespass to chattel regarding access to a computer system, a plaintiff must show: "(1) the defendant intentionally and without authorization interfered with plaintiff's possessory interest in the computer system; and (2) the defendant's unauthorized use proximately resulted in damage to plaintiff."[1]

In *Ebay Inc. v. Bidder's Edge, Inc.*, one of the leading cases on this type of claim, the court found that Bidder's Edge's ("BE") use of web crawlers on eBay's website amounted to trespass to chattel. The court first determined that BE's use of robots, which accessed eBay's site approximately 100,000 times a day, exceeded the "scope of consent" granted by eBay even though the website was publicly accessible.[2] The court rejected the argument that eBay was required to present evidence of "substantial interference" with possession.[3] Instead, it was sufficient to show that BE's

conduct was at least "intermeddling with the use of another's personal property."[4] The court further noted that although it felt "some uncertainty as to the precise level of possessory interference required to constitute an intermeddling, there does not appear to be any dispute that eBay can show that BE's conduct amounts to use of eBay's computer system."[5]

The court then addressed the damages element of the claim, holding that "BE's activities have diminished the quality or value of eBay's computer systems [by consuming] at least a portion of plaintiff's bandwidth and server capacity."[6] The court went on to note that eBay's server and its capacity were personal property, and that BE's searches used a portion of this property, thereby depriving eBay of the ability to use that portion, even if those searches used only a small amount of eBay's computer system capacity.[7] The court further observed that eBay could be damaged in the future if this conduct was left unchecked, since other competitors could implement similar programs, which, when aggregated, would cause serious impairment to eBay's system.[8]

The *eBay* holding, however, may be more limited than it appears. First, the trespass to chattels action gives standing only to possessors of chattels, for the action protects the possessor's ability to exclude others from using the chattels.[9] The eBay court found that eBay's *computer servers* were chattels, and impairment of the server capacity constituted an interference with eBay's possessory interest in the servers.[10] Following this line of reasoning, the only proper plaintiff in a trespass to chattels case would potentially be the server's owner.[11] This claim would thus more likely protect only corporate plaintiffs, like eBay, and others large enough to own their own servers. A plaintiff whose website resides on the server of a commercially available internet service provider may not have a protectable interest as a possessor of chattels.[12]

Other courts have adopted an even more stringent requirement for establishing trespass to chattel than that of the *eBay* court.[13] The court in *Ticketmaster Corp. v. Tickets.com* held that "unless there is actual dispossession of the chattel for a substantial time (not present here), the elements of the tort have not been made out. Since the spider does not cause physical injury to the chattel, there must be some evidence that the use or utility of the computer (or computer network) being 'spiderized' is adversely affected by the use of the spider."[14] The court also rejected the argument that plaintiff's efforts to stop the spider, and the value of the information taken by the spider, provided a basis for damage, noting that "neither of these items shows damage to the computers or their operation."[15]

Complicating the issue of damages even further, in 2003 the California Supreme Court found that certain types of service providers might be able to raise loss of business reputation or goodwill to satisfy the harm element of the claim, but others cannot. In *Intel Corp. v. Hamidi*, the Court found that a complaining party must demonstrate "an injury to its personal property, or to its legal interest in that property."[16] Consequently, evidence that Intel's employees' time was occupied by reading or attempting to block the defendant's offending e-mail messages did not satisfy the harm element of trespass to chattels. While these costs represented real and measurable losses to Intel, they did not represent any decrease in the value of Intel's property interests. Furthermore, while other federal cases had arguably opened the door for a trespass to chattels claim based on loss of business reputation or goodwill,[17] the Court restricted those holdings to their facts. The Court noted that while an Internet Service Provider ("ISP") may have an argument that loss of goodwill constitutes harm to its legally protected interests in its chattels, Intel's claimed injury had less connection to its personal property than an ISP.[18] Presumably, this is because unlike an ISP, Intel's goodwill is not a function of the quality of the internet services it can provide.

These contrasting and intricate opinions demonstrate the complexities of the trespass to chattel theory of liability and the importance of recognizing what types of possessory interests and damages a court will consider in evaluating such a claim.

BREACH OF CONTRACT

Most commercial websites have terms of use governing how the website and its content can be used (and if they don't, they should). A website's terms of use can be, under certain circumstances, enforced against a website visitor who uses the website in violation of those terms. Breach of contract liability rests on the theory that a contract was formed between the website provider and the visitor, whereby the visitor accesses the website and in exchange agrees to be bound by the provider's terms of use. Fundamental to the contract formation, however, is that the website user had actual or constructive knowledge of the website's terms and conditions and assented to those terms.[19] There are two general methods by which terms of use are presented to a website visitor, and the chosen method affects the website provider's ability to enforce those terms against a violator. These two methods are referred to as "browsewrap" and "clickwrap" agreements.

A clickwrap agreement involves the active presentation of the terms of use to the visitor on the screen, with the requirement that the visitor click through to continue using the website (this is often accomplished through an "I agree" button or clicking on a box next to the words "I agree to these terms and conditions" or similar). Although there is some discussion among the courts about the length, density and legal language used in the terms of use presented to the visitor, clickwrap agreements are generally considered to be enforceable against visitors, since the visitor must give affirmative assent to the terms before being granted further access to the website.[20]

A browsewrap agreement is less direct than a clickwrap agreement, requiring no affirmative conduct by the visitor. Generally, terms of use are provided on the website via a hyperlink (usually located at the bottom of the main webpage), which directs the visitor to another page or remote location to view the terms of use. The user is not required to click on anything to accept the terms of use before being granted access to the contents of the site. Because there is no guarantee that the visitor viewed, or has notice or knowledge of, the terms of use, courts are much more reluctant to enforce browsewrap agreements.[21]

A visitor who frequently accesses a particular site, however, is more likely to be found to have knowledge of the site's terms of use, such that a browsewrap agreement could be enforced against that visitor. Moreover, courts will generally hold a sophisticated user that builds a business on using information from third-party sites to a higher standard than a non-business/consumer user in terms of notice and enforcement of browsewrap terms of use.[22]

The use of an automated process, such as a "bot" or "spider" program, to crawl third party websites will often be a direct violation of a site's terms of use because websites typically prohibit any automated program that operates more quickly than

a human performing the same tasks, as these programs may put a strain on the website's resources. Moreover, courts have found that the use of an automated program is sufficient to impute constructive notice of terms of use.[23] In addition, if the bot is programmed to defeat security measures to reach the desired content, this significantly increases the chance of liability.

TRADEMARK INFRINGEMENT

The federal Lanham Act, 15 U.S.C. § 1051 *et seq.*, provides broad protection for trademark holders, and prohibits any use of a trademark that may cause confusion or deception as to the origin of the product, or as to a relationship, affiliation or endorsement between the trademark owner and the alleged infringer.[24] Because scrapers often provide links to third-party content on their own websites, there is an inherent risk of infringing (whether knowingly or inadvertently) third-party trademarks by confusing consumers as to the source of the marks being used. Moreover, even providing a link to third-party content contained within the third-party websites (referred to as "deep linking") can create confusion and disadvantage if the user is not required to visit the third-party site, which itself may contain revenue-generating advertising or important terms and conditions.

Web scrapers will often attempt to defend a claim of trademark infringement by arguing that any trademark use is nominative fair use. Under nominative fair use, there is no trademark violation if the trademark owner's mark is used in a non-confusing way to identify the trademark owner's goods or services.[25] Such use does not imply sponsorship or endorsement, but instead uses the mark in a competitive or comparative manner that is not likely to confuse customers.[26] One of the requirements of a nominative fair use defense, however, is that "the user must do nothing that would, in conjunction with the mark, suggest sponsorship or endorsement by the trademark holder."[27] Whether a web scraper's use of a trademark is in violation of the Lanham Act is a fact-specific and nuanced determination, but trademark owners need to remain vigilant in protecting their rights in the mark.[28] Consequently, owners should carefully monitor the use of their protected marks on third party websites.

THE COMPUTER FRAUD AND ABUSE ACT

As with a trespass to chattels claim, if a web scraping program accesses a third-party site without authorization, copies data or stores it for later use, distributes the website or uses it for commercial use, the scraper may run afoul of the Computer Fraud and Abuse Act, 18 U.S.C. § 1030 ("CFAA"). Specifically, the CFAA provides criminal punishment and/or fines to

> Whoever. . . intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains. . . information from any protected computer; [or] . . . knowingly causes the transmission of a program, information, code, or command, and as a result of such conduct, intentionally causes damage [defined as any impairment to the integrity or availability of data, a program, a system, or information] without authorization, to a protected computer [resulting in a] loss to one or more persons during any 1-year period . . . aggregating at least $5,000 in value; [or] intentionally accesses a protected computer without authorization, and as a result of such conduct, causes damage [resulting in a] loss to 1 or more persons during any 1-year period . . . aggregating at least $5,000 in value . . ..[29]

There are a couple of important points to emphasize in relation to liability under the CFAA. First, the CFAA applies to all companies, whether for profit or non-profit and all computers as long as they are connected to the internet. Second, while the CFAA is primarily a criminal statute, it also provides that any person who suffers damages by reason of violation of the statute can bring a civil action against the violator to recover compensatory damages and injunctive relief.[30] Third, like (or unlike, as the case may be) the trespass to chattels claim discussed above, loss of business or goodwill are appropriate forms of loss or damage under the statute.[31] Finally, depending on which circuit governs the jurisdiction where the case is filed, the lack of authorization can be alleged and proven based on a breach of the agency relationship; a breach of an employment contract such as a violation of company rule; or a use of the computer that exceeds expected norms of authorized use.

While undoubtedly a useful tool for website owners to protect their sites, there is no bright-line test for when a scraping activity violates the CFAA as activity that "exceeds expected norms of authorized use." Many courts have found that use of a website in violation of its stated terms of use constitutes unauthorized use for purposes of the statute.[32] Courts have also suggested that a clear statement by a website provider that scraping is unauthorized will give rise to cause of action under the CFAA.[33] However, many of these decisions have been criticized in academic and technology circles as perverting the CFAA's intent to curb hacking by affording website owners a method for obtaining absolute control over access to and use of information they have chosen to post on their publicly available sites.[34]

In one recent case, a federal court in the Eastern District of Virginia emphasized the particularity of the CFAA. In *Cvent, Inc. v. Eventbrite, Inc.*, Cvent argued that a competing event company, Eventbrite, was scraping valuable data from Cvent's online database for their own commercial use in violation of the CFAA.[35] In dismissing the claim, the court noted the distinction between unauthorized *use* of data and unauthorized *access* to data because only unauthorized access constitutes a violation of the CFAA.[36] Cvent's database was publicly available on its website and could be accessed without any password or login information. As the court put it, "the entire world was given unimpeded access" to Cvent's database and Eventbrite had therefore not violated the CFAA when it scraped that information.[37] Cvent's argument that they had limited Eventbrite's access by virtue of a browsewrap agreement containing its terms of use was of no avail because the browsewrap was not prominently displayed and thus no reasonable user could be expected to notice it.[38]

The cases described above offer important lessons for website owners concerned with scraping. Owners should consider particular steps such as password protection, clickwrap agreements or obvious alerts to the user that the site is governed by an end user license agreement to protect their interests.

CONCLUSION

While the theories of liability are still developing and court decisions are not uniform, there is a pattern emerging that the courts are prepared to protect proprietary content on commercial websites from uses

which are detrimental to the site owners. The degree of protection, however, for such content is not settled and will depend on the type of access made by the scraper, the terms of use of the site being scraped, the amount of information scraped and the type of adverse effects on the owner's system. To better protect their data, website owners should consider adding or revising appropriate terms of use to their sites and clearly specifying how their website content can be displayed, accessed and used by site visitors, as well as prohibitions on such access and use. In addition, website owners desiring to bind their users to terms of use should consider requiring assent by use of a clickwrap or other means to ensure that the terms of use and modifications to such terms are available to the user prior to use and are conspicuously placed on the site.

ENDNOTES

1. *eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058, 1069 (N.D. Cal. 2000).
2. *Id*. at 1070.
3. *Id.*
4. *Id.*
5. *Id.*
6. *Id.* at 1071
7. *Id.*
8. *Id.*
9. A person who is in possession of chattel is one who has physical control of it. Restatement (Second) of Torts 216 (1965).
10. *eBay*, 100 F. Supp. 2d at 1070-71 (holding BE's trespass occurred on eBay's privately-owned servers, not on its publicly-accessible website).
11. The eBay court suggested only that the "owner" of personal property may recover actual damages suffered from impairment of the property or the loss of its use. *Id.* at 1065. It is unclear whether an individual subscriber to an ISP would have, by virtue of the subscription, a legal ownership interest in the ISP's server sufficient to establish a trespass to chattels action.
12. If the impairment of server capacity is the measure of harm, as the eBay court found, then it would appear that the only proper plaintiff in a trespass to chattels case would be the owner of the impaired server.
13. *Ticketmaster Corp. v. Tickets.com*, No. CV99-7654-HLH, 2003 U.S. Dist. LEXIS 6483 (C.D. Cal. Mar. 6, 2003).
14. *Id.* at 12.
15. *Id.* at 13.
16. *Intel Corp. v. Hamidi*, 1 Cal. Rptr. 3d 32, 47 (2003).
17. *See, e.g., Compuserve, Inc. v. Cyber Promotions, Inc.*, 962 F. Supp. 1015 (S.D. Ohio 1997).
18. *Intel*, 1 Cal. Rptr. 3d at 45.
19. *See, e.g., Southwest Airlines Co. v. BoardFirst, L.L.C.*, No. 06-CV-0891-B, 2007 U.S. Dist. LEXIS 96230, 15-16 (N.D. Tex. Sept. 12, 2007) ("the validity of a browsewrap license turns on whether a website user has actual or constructive knowledge of a site's terms and conditions prior to using the site"); Mark A. Lemley, *Terms of Use*, 91 Minn. L. Rev. 459, 477 (Dec. 2006) ("Courts may be willing to overlook the utter absence of assent only when there are reasons to believe that the defendant is aware of the plaintiff's terms").
20. *See, e.g., Specht v. Netscape Communs. Corp.*, 306 F.3d 17, 22 (2d Cir. 2002) ("clicking on a webpage's clickwrap button after receiving notice of the existence of license terms has been held by some courts to manifest an Internet user's assent to terms governing the use of downloadable intangible software").
21. *Id.* at 32 (no basis to impute knowledge of terms of use where user would not have seen terms without scrolling down computer screen, but doing so was unnecessary to complete download).
22. *See, e.g., Register.com v. Verio, Inc.*, 126 F. Supp. 2d 238 (S.D.N.Y. 2000), *aff'd*, 356 F.3d 393 (2d Cir. 2004) (the court was willing to enforce a browsewrap agreement against a competitor who routinely accesses a competing website to gather data or content).
23. *See, e.g., Register.com*, 356 F.3d at 401-02.
24. *See* 15 U.S.C. § 1114(1)(a) (prohibiting use of "any reproduction, counterfeit, copy, or colorable imitation of a registered mark" that is "likely to cause confusion, or to cause mistake, or to deceive"); 15 U.S.C. § 1125(a) (prohibiting use of "any word, term, name, symbol, or device, or any combination thereof, or any false designation of origin" that is "likely to cause confusion, or to cause mistake, or to deceive as to the affiliation, connection, or association").
25. *See New Kids on the Block v. News America Pub., Inc.*, 971 F.2d 302 (9th Cir. 1992).
26. *Id.*
27. *New Kids*, 971 F.2d at 308.
28. *Amstar Corp. v. Domino's Pizza, Inc.*, 615 F.2d 252, 265 (5th Cir. 1980) (narrowing of protection of mark where plaintiff failed to be vigilant in protecting its rights in the mark).
29. 18 U.S.C. § 1030.
30. 18 U.S.C. 1030(g).
31. *See e.g., Creative Computing v. Getloaded.com, LLC*, 386 F.3d 930, 935 (9th Cir. 2004).
32. *See e.g., EF Cultural Travel BV v. Explorica, Inc.*, 274 F.3d 577, 582-82 (1st Cir. 2001) (finding that defendant's use of a computerized "scraper" to glean information from plaintiff's website likely exceeded authorized access where such use at least implicitly violated a confidentiality agreement); *Southwest Airlines v. Farechase, Inc.*, 318 F. Supp. 2d 435, 439-40 (N.D. Tex. 2004) (finding that Southwest sufficiently stated CFAA claim where Southwest had directly informed the defendant that its scraping of southwest.com was unauthorized).
33. *See e.g., EF Cultural Travel BV v. Zefer Corp.*, 318 F.3d 58, 64 (1st Cir. 2003) ("[W]ith rare exceptions, public website providers ought to say just what non-password protected access they purport to forbid").
34. *See e.g.*, Christine D. Galbraith, *Access Denied: Improper Use of the Computer Fraud and Abuse Act to Control Information on Publicly Accessible Internet Websites*, 63 MD. L. REV. 320, 368 (2004); Mark A. Lemley, *Place and Cyberspace*, 91 CAL. L. REV. 521, 528 (March 2003) ("An even more serious problem is the judicial application of the [CFAA], which was designed to punish malicious hackers, to make it illegal -- indeed, criminal -- to seek information from a publicly available website if doing so would violate the terms of a 'browsewrap' license.").
35. 2010 U.S. Dist. LEXIS 96354 (E.D. Va. Sept. 14, 2010).
36. *Id* at 8-9.
37. *Id* at 13.
38. *Id* at 9.