Bloomberg BNA

Electronic Commerce & Law Report™

VOL. 18, NO. 34 AUGUST 28, 2013

CONTRACTS

The authors examine the legal impacts of utilizing web crawlers and scrapers in an era of intensive ''big data'' demands, including issues relating to copyright, contract, trespass, and the Computer Fraud and Abuse Act.

Use of Online Data in the Big Data Era: Legal Issues Raised by the Use of Web Crawling and Scraping Tools For Analytics Purposes

By Jim Snell and Derek Care

In 2010, Pete Warden, a software engineer living in Colorado, developed a software program to ''crawl'' publicly accessible Facebook pages and ''scrape'' (*i.e.*, collect) information relating to Facebook's members. Within hours of deploying his software, the application had visited approximately 500 million pages and collected information related to approximately 220 million Facebook users – including users' names, location information, friends and interests. Using this dataset, which Mr. Warden offered to release in anonymized form for research purposes, he created a graphical analysis of the regional and relationship patterns among Facebook's members. The cost of this exercise: about $100. The results: more than 500,000 visits to Mr. Warden's website, national media coverage, and cease-and-desist warnings from Facebook, which perceived Mr. Warden's collection of data from its webpages as a violation of its terms of use prohibiting automated access to the website without the company's permission. Ultimately, in order to avoid a potential legal dispute, Mr. Warden abandoned his plan to release the information he collected, and agreed to delete all copies of the dataset.[1] Summing up his experience, he later quipped, ''Big data? Cheap. Lawyers? Not so much.''[2]

Jim Snell is co-chair of Bingham's Intellectual Property Group and co-chair of the firm's Privacy and Security Group. He represents clients in complex commercial matters, including patent litigation, Internet and privacy issues, trade secret matters and matters involving unfair competition claims under California Business and Professions Code section 17200.

Derek Care is counsel in Bingham's New York office, where he is a member of the firm's Intellectual Property, Privacy and Security, and Entertainment, Media and Communications groups. He represents clients in a wide range of complex commercial matters, including trademark, copyright, Internet and privacy-related matters.

Automated Web Content Gatherers

The use of web crawlers, scrapers and others automated tools for gathering online content has long been a feature of Internet (to the extent ''long'' can be used

[1] *See* Pete Warden, *How I Got Sued by Facebook*, PeteSearch (Apr. 5, 2010), http://petewarden.typepad.com/searchbrowser/2010/04/how-i-got-sued-by-facebook.html; Jim Giles, *Data Sifted from Facebook Wiped after Legal Threats*, New Scientist (Mar. 31, 2010), http://www.newscientist.com/article/dn18721-data-sifted-from-facebook-wiped-after-legal-threats.html; Leon Erlanger, *Big Data Runs Afoul of Big Lawyers*, InfoWorld (Mar. 28, 2011), http://www.infoworld.com/t/big-data/big-data-runs-afoul-big-lawyers-854; Janko Roettgers, *Data Science Toolkit Brings Big Data Analysis to the People*, GigaOM.com (Mar. 23, 2011, 1:27 PM), http://gigaom.com/2011/03/23/pete-warden-openheatmap-data-science-toolkit.

[2] *See* Roettgers, *supra* note 1.

 ISSN 1098-5190

to describe the history of the Internet). For example, searches engines use web crawling ''bots'' or ''spiders'' to continuously visit billions of webpages to create relevant and accurate search results, and the Internet Archive – a non-profit digital library that archives historical versions of publicly accessible webpages – has since 1996 used web crawling tools to create a historical record of the Internet comprising 10 quadrillion bytes of data. Others have used similar tools to offer services that compete with or complement the offerings of the scraped websites – including uses of these tools to aggregate news content, and to monitor and facilitate purchases of airlines and concert tickets (with or without the permission or involvement of the scraped website). As Mr. Warden's experience suggests, the use of these tools pit the interests of website owners in protecting, controlling and profiting from the content they provide against the interests of others who seek to gather and use that content for other purposes (be they harmful, helpful or irrelevant to the website owner). Not surprisingly, the use of these tools has spurred litigation under a variety of theories, including copyright infringement, breach of contract (*e.g.*, website terms of use), ''hot news'' misappropriation, trespass to chattels, and criminal statutes prohibiting unauthorized access to a computer system or website.

With the advent of Big Data – the increasingly widespread practice of using advanced data analytics to identify trends and patterns in extremely large datasets collected from a variety of sources – the potential applications for scraped data, and the benefits associated with analysis of that data, have increased exponentially. Whereas past cases involving unauthorized web crawling and scraping often involved simple copying and republication of website content in direct competition with the scraped website, the growing use of advanced data analytics is giving rise to instances where the connection between the data analytics service and the scraped website is attenuated and not directly competitive. Nevertheless, the online content of websites that may be scraped is among such businesses most valuable data, and great lengths are understandably taken to protect such content.

Given both the tremendous value and Big Data-driven demand for Internet-based information, and the relative ease by which such information can be compiled using automated data collection tools such as that deployed by Mr. Warden, it is likely that future cases relating to web crawling and scraping will focus on the legal issues raised by automated data gathering for analytics purposes – and what theories a website owner may exercise to protect any factual data so collected and what theories a data collector may use to justify such collection. Few courts, however, have directly addressed the legal issues raised by Big Data or the collection of data for related purposes, leaving uncertain the legal environment faced by website owners wishing to protect the data on their websites, and those who would gather such data for analytics purposes. Without taking sides – and while recognizing that the legal landscape relating to the Internet is constantly evolving, with previously challenged technologies such as search engines now recognized as nearly *per se* legitimate while others such as peer-to-peer networks have continually been subject to scrutiny – this article seeks to outline the legal issues such parties may face. In doing so, this article will consider the legal theories that have been applied in prior cases relating to the use of web crawling and scraping tools in other contexts, and will identify issues relating to whether claims under these theories are likely to succeed in connection with disputes relating to automated data collection for Big Data and analytics purposes.

Legal Theories Related to Automated Online Data Collection

A. Copyright Infringement.

The Copyright Act protects original expressions that are fixed in a tangible medium, including mediums such as computer memory or a web server.[3] These protections extend not only to original expressions such as images contained on a website, but also the underlying code that enables the display of any content on the website – including facts displayed on a website that are not otherwise entitled to copyright protection. Accordingly, because web crawling and scraping tools generally index information on a targeted webpage regardless of whether the tool seeks to obtain copyrighted content or unprotected facts[4], courts have recognized claims for copyright infringement in connection with the use of web crawling and scraping tools.[5]

Because some courts have recognized that such activities may infringe a website owner's copyrights, the focus in such cases is generally on whether the web crawling or scraping at issue is a fair use of the copyrighted content. For example, in *Kelly v. Arriba Soft*

[3] *See, e.g.*, *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1160 (9th Cir. 2007) (discussing copyright protections available to works ''embodied (i.e., stored) in a computer's server . . . or hard disk, or other storage device. . . .'').

[4] *See Ticketmaster Corp. v. Tickets.com, Inc.*, No. CV997654HLHVBKX, 2003 BL 2425, at *2-3 (C.D. Cal. Mar. 7, 2003) (discussing technical operation of web crawling and scraping tool).

[5] *See, e.g.*, *Facebook, Inc. v. Power Ventures, Inc.*, No. C 08-05780 JF (RS), 2009 BL 288736, at *4 (N.D. Cal. May 11, 2009) (denying motion to dismiss where allegations that defendant momentarily created ''cache'' copies of Facebook's webpages, including the protected elements thereon, sufficiently stated a claim for copyright violation); *see also Tickets.com*, 2003 BL 2425, at *7 (granting summary judgment dismissing Ticketmaster's copyright claim because, though defendant momentarily copied the protected elements on Ticketmaster's website in order to extract non-copyrightable factual information thereon, such copying was a fair use of Ticketmaster's protected content). For a further discussion of fair use as a defense to copyright claims relating to momentary copying of protected content for the purpose of extracting factual information from a website, *see* pp. 7-9, *supra*.

Corp., the defendant search engine conceded that its display of low-resolution ''thumbnail'' copies of high-resolution photographs constituted reproduction of those photographs, but argued that such display was a transformative, fair use of the copied photographs. The Ninth Circuit agreed – holding that the search engine's display of low-resolution photographs to facilitate the general public's access to information on the Internet was highly transformative of, and did not provide a substitute for, the plaintiff's high-resolution photographs whose purpose was primarily artistic.[6] Notably, the fact that such use was for a commercial purpose did not bar the court's finding that the search engine made a fair use of plaintiff's copyrighted photograph.[7] In contrast, in *Associated Press v. Meltwater Holdings U.S., Inc.*, the court found that an online news aggregator that provided its subscribers with nearly 500-character excerpts of copyrighted articles scraped from the website's of the Associate Press's licensees did not engage in a fair use of those articles. The court distinguished the news aggregator's services from those at issue in *Kelly* on the grounds that the news aggregator did not facilitate the general public's access to information on the Internet, but instead only provided word-for-word excerpts of the copied articles to the aggregator's paying customers without transforming that content in any way.[8] The court further held that the aggregator's use of that content to generate analytics relating to the online news sources it covered, while potentially transformative in and of itself, did not render the aggregator's excerpting transformative insofar as the analytics and excerpting were separate and distinct services.[9]

While even incidental reproduction of copyrighted webpage material may give rise to copyright liability, courts have also recognized that such reproduction may constitute a fair use of the protected content. For example, in *Ticketmaster Corp. v. Tickets.com, Inc.*, the defendant argued that the momentary copying of Ticketmaster's webpages by its spiders for the purpose of extracting factual information concerning concert times, ticket prices, and venues that defendant then posted to its website constituted a fair use. The court agreed. In so finding, the court emphasized that the copying was momentary, the effect on the market value of the copyrighted material was ''nil'', and that the ''amount and substantiality'' of the material used was negligible insofar as defendant did not reproduce the copyrighted material on its webpage. Further, the court observed that the central purpose of the Copyright Act – *i.e.*, ''to secure a fair return for an author's creative labor and to stimulate artistic creativity for the general good'' – would not be served by restricting defendant from momentarily copying Ticketmaster's webpages for the purpose of obtaining non-protected, factual information.[10]

In addition to the fair use defense, courts have also considered whether a plaintiff's copyright claims are subject to implied license or estoppel defenses based on its failure to deploy the ''robots.txt'' protocol to deter unwanted web crawling or scraping. The robots.txt protocol is industry-standard programming language that a website may deploy to instruct cooperating web crawlers generally, or certain web crawlers specifically, to *voluntarily* refrain from accessing all or part of the website.[11] In *Parker v. Yahoo, Inc.*, the court held that the plaintiff's failure to deploy the protocol granted Yahoo an implied license to create cache copies of his website where plaintiff was aware that Yahoo – which has a policy of not creating cache copies of websites that deploy the protocol – would do so in the absence of the protocol.[12] Conversely, in *Meltwater*, the court rejected the defendants' implied license and estoppel defenses based on the Associated Press's purported failure to require its licensees to deploy the protocol. The court distinguished *Parker* on several grounds, including that the defendants reserved the right to ignore the protocol if deployed. The court further emphasized that the defendants' arguments, if accepted, would shift the burden of preventing infringement to the copyright owner, and threatened the ''openness of the Internet'' by forcing copyright owners to choose between deploying the protocol and deterring all web crawlers (including search engines which may help users locate the website), and refraining from doing so and losing the right to prevent unauthorized use of its protected content.[13]

With respect to future cases involving use of scraped content for analytics purposes, courts are likely to follow a similar analysis driven by the facts of the specific case. Issues regarding whether the copying is momentary, whether the information extracted is factual, the effect on the market value of the copyrighted material, and the amount and substantiality of the material used are likely to be key issues in these cases. Courts are further likely to focus on whether the object of the Copyright Act – ''to secure a fair return for an author's creative labor and to stimulate artistic creativity for the general good'' – would be served by prohibiting the

[6] 336 F.3d 811, 818-22 (9th Cir. 2003); *see alsoPerfect 10*, 508 F.3d at 1154-55, 1165, 1168 (holding that Google's display of low-resolution thumbnail images in its search results was a fair use of plaintiff's copyrighted photographs).

[7] *Kelly*, 336 F.3d at 818.

[8] No. 12 Civ. 1087(DLC), 2013 BL 74727, at *4, *13-22 (S.D.N.Y. Mar. 21, 2013); *see id.* at *12 (holding that the news aggregator's ''use[] [of] its computer programs to automatically capture and republish designated segments of text from news articles, without adding any commentary or insight in its News Reports'' constitutes ''undiluted use'' of the Associated Press's copyrighted articles).

[9] *Id.* at *17.

[10] *See* No. CV997654HLHVBKX, 2003 BL 2425, at *6, *7-8 (C.D. Cal. Mar. 7, 2003) (discussing the four fair use factors set forth in 17 U.S.C. § 107). *See also Nautical Solutions Marketing, Inc. v. Boats.com*, No. 8:02–CV–760–T–23TGW, 2004 BL 3401, at *2 (M.D. Fla. Apr. 1, 2004) (''Boat Rover's momentary copying of Yachtworld's public web pages in order to extract from yacht listings facts unprotected by copyright law constitutes a fair use and thus 'is not an infringement of copyright.' '') (quoting 17 U.S.C. § 107).

[11] Use of, and compliance with, the protocol is voluntary, not mandatory; while some website and content owners have promoted more robust versions of the protocol, and even legislation that would make compliance with the protocol mandatory, such efforts have not been successful.

[12] No. 07 Civ. 2757, 2008 BL 215777, at *4 (E.D. Pa. Sept. 25, 2008).

[13] *Meltwater*, 2013 BL 74727, at *25 (''It is fair to assume that most Internet users (and many owners of websites) would like crawlers employed by search engines to visit as many websites as possible, to include those websites in their search results, and thereby to direct viewers to a vast array of sites. Adopting Meltwater's position would require websites concerned about improper copying to signal crawlers that they are not welcome.'').

challenged conduct. Courts are also likely to consider, in the context of defenses to copyright claims, the specific circumstances relating to a website's deployment of the robots.txt protocol, including whether the defendant has a practice or policy of complying with the protocol if deployed.

B. Breach of Contract.

Most commercial websites contain terms of use that provide that access and/or use of the website is premised on the user's agreement to such terms.[14] A claim sometimes made in cases regarding web crawling or scraping is that the defendant violated the terms of use by crawling and scraping content. While these cases have explored somewhat novel uses of technology, they often turn on fundamental issues of contract[15] – including whether the targeted website's terms of use are enforceable as against the defendant, whether the conduct complained of violates those terms, and whether any such violation causes any compensable damages. These cases suggests that use of such tools to gather data may give rise to a claim for breach of contract, while also demonstrating the potential hurdles to prevailing on such claims. These issues are discussed in turn.

1. Enforceability of Website Terms of Use.

As is the general rule with any contract, a website's terms of use will generally be deemed enforceable if mutually agreed to by the parties. In determining whether such mutual agreement exists, courts look to whether the terms of use constitute a ''clickwrap'' agreement – which typically require that a visitor indicate her agreement by clicking an ''I accept'' icon before accessing the website – or a ''browsewrap'' agreement – pursuant to which the user is provided with notice of the website's terms of use, and informed that use of the website constitutes agreement to those terms.[16] Clickwrap agreements, because they require a user to formally indicate his knowledge and awareness of the terms of use, are generally found enforceable.[17] Browsewrap agreements have also generally been found enforceable where the defendant has actual knowledge of the terms of use or constructive knowledge of such terms.[18] Actual knowledge is sometimes demonstrated by evidence that a defendant was advised of its violations of the terms of use via a cease-and-desist letter from plaintiff.[19] Constructive knowledge is sometimes found where a website's terms of use are prominently or conspicuously displayed on the website, such as where a hyperlink to those terms is underlined and set forth in distinctively colored text.[20]

Regardless of whether a website's terms of use are clickwrap or browsewrap, the defendant's failure to read those terms is generally found irrelevant to the enforceability of its terms.[21] One court disregarded arguments that awareness of a website's terms of use could not be imputed to a party who accessed that website using a web crawling or scraping tool that is unable to detect, let alone agree, to such terms.[22] Similarly, one court imputed knowledge of a website's terms of use to a defendant who had repeatedly accessed that website using such tools.[23] Nevertheless, these cases are, again, intensely factually driven, and courts have also declined to enforce terms of use where a plaintiff has failed to sufficiently establish that the defendant knew or should have known of those terms (*e.g.*, because the terms are inconspicuous), even where the defendant repeatedly accessed a website using web crawling and scraping tools.[24]

---

[14] Mark A. Lemley, *Terms of Use*, 91 Minn. L. Rev. 459, 460 (Dec. 2006) (defining terms of use as the agreements that ''control (or purport to control) the circumstances under which . . . visitors to a public Web site can make use of that . . . site'').

[15] *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 403 (2d Cir. 2004) (''While . . . the Internet has exposed courts to many new situations, it has not fundamentally changed the principles of contract.'').

[16] *Hines v. Overstock.com, Inc.*, 668 F. Supp. 2d 362, 366 (E.D.N.Y. 2009) (''On the internet, the primary means of forming a contract are the so-called ''clickwrap'' (or ''click-through'') agreements, in which website users typically click an ''I agree'' box after being presented with a list of terms and conditions of use, and the ''browsewrap'' agreements, where website terms and conditions of use are posted on the website typically as a hyperlink at the bottom of the screen.'') (citing *Register.com*, 356 F.3d at 403).

[17] *See, e.g.*, *Specht v. Netscape Comm'cns Corp.*, 306 F.3d 17, 22 n.4 (2d Cir. 2002) (''[C]licking on a webpage's clickwrap button after receiving notice of the existence of license terms has been held . . . to manifest an Internet user's assent to terms governing the use of downloadable intangible software.'') (citing *Hotmail Corp. v. Van$ Money Pie Inc.*, 47 U.S.P.Q.2d 1020, 1025 (N.D. Cal. 1998) (finding clickwrap terms of use enforceable)); *see also* Lemley, *supra* note 13, at 466 (''[E]very court to consider the issue has held clickwrap licenses enforceable.'').

[18] *Register.com*, 356 F.3d at 401-04 (constructive knowledge); *Southwest Airlines Co. v. BoardFirst, L.L.C.*, No. 3:06-CV-0891-B, 2007 BL 114340, at *4 (N.D. Tex. Sept. 12, 2007) (actual knowledge); *Cairo, Inc. v. Crossmedia Servs., Inc.*, No. C 04-04825 JW, 2005 BL 9669, at *4-5 (N.D. Cal. Apr. 1, 2005) (constructive knowledge).

[19] *See Southwest Airlines*, 2007 BL 114340, at *4.

[20] *See PDC Labs. v. Hach Co.*, No. 09-1110, 2009 BL 293520, at *3 (C.D. Ill. Aug. 25, 2009) (finding that terms and conditions posted on website were sufficiently conspicuous that knowledge of their terms could be imputed to plaintiff where those terms were ''hyperlinked on three separate pages . . . in underlined, blue, contrasting text.'').

[21] *See id.* at *2-3 (''Whether a party has actually read terms and conditions of sale documents does not affect their . . . enforceability. . . . Though PDC does not state whether or not they read the Terms, it is inconsequential to . . . PDC's [assent to those terms].'') (citing *Druyan v. Jagger*, 508 F. Supp. 2d 228, 237 (S.D.N.Y. 2007) (holding that plaintiff was bound to online terms of use regardless of whether she actually read them)).

[22] *Internet Archive v. Shell*, 505 F. Supp. 2d 755, 765 (D. Colo. 2007) (denying motion to dismiss breach of contract claim based on purported violation of website's terms of use, and rejecting defendant's argument that plaintiff failed to adequately allege that the terms were enforceable insofar as she did not allege that any human employee of defendant accessed plaintiff's website).

[23] *See, e.g.*, *Cairo,* 2005 BL 9669, at *5 (''Cairo's repeated and automated use of CMS's web pages can form the basis of imputing knowledge to Cairo of the terms on which CMS's services were offered . . . .'') (citing *Register.com*, 356 F.3d at 401-02 (imputing knowledge of web site's terms of use to repeated user of Register.com's database)).

[24] *See Cvent, Inc. v. EventBrite, Inc.*, 739 F. Supp. 2d 927, 937 (E.D. Va. 2010) (plaintiff's allegation that the terms of use of its website were readily available for review on its website was insufficient to plausibly establish that defendant was aware of those terms where website did not require that users manifest awareness of or consent to those terms, and where terms themselves could only be reached via a ''small link[] . . . buried at the bottom of the first page'' of plaintiff's website);

   ISSN 1098-5190

Issues regarding enforceability of contract are likely to continue to be an issue addressed by courts in this area, with content providers citing clickwrap agreements and actual knowledge of terms, and those using crawling and scraping tools arguing a lack of mutual assent to such terms.

2. Terms of Use That May Prohibit Automated Data Collection.

The terms of use for websites frequently include clauses prohibiting access or use of the website by web crawlers, scrapers or other robots, including for purposes of data collection. Courts have recognized causes of action for breaches of contract based on the use of web crawling or scraping tools in violation of such provisions.[25]

Also common are terms of use that limit visitors to personal and/or non-commercial use of a website. For example, in *Southwest Airlines Co. v. BoardFirst, LLC*, the plaintiff airline alleged that the defendant violated its terms of use restricting access to Southwest's website for ''personal, non-commercial purposes'' by offering a commercial service that helped Southwest's customers take advantage of the company's ''open'' seating policy and check-in process to obtain priority seating in the front of the plane. The court granted Southwest's motion for summary judgment on its breach of contract claim, finding that the defendant's conduct directly contravened Southwest's prohibition on commercial uses of Southwest's website.[26]

Cases addressing the purported violations of these terms tend to hinge on the precise language of the contractual provisions at issue, and the scope of the agreement between the parties that can be ascertained from that language. Thus, for example, in *Southwest*, the court rejected defendant's argument that Southwest's terms of use were too ambiguous to be enforced against defendant where those terms specifically prohibited use of the website ''for the purpose of checking [c]ustomers in online or attempting to obtain for them a boarding pass in any certain boarding group.'' Defendant's services, which helped Southwest's customers obtain priority seating, fell ''within the heart of this proscription.''[27] In contrast, in *TrueBeginnings, LLC v. Spark Network Servs., Inc.*, the court found that the defendant did not violate the terms of service of plaintiff's dating website – which limited use of the ''website and related services'' to a visitor's ''sole, personal use'' – by visiting the website to obtain evidence for use in a patent infringement action against plaintiff. In so holding, the court analyzed the entirety of plaintiff's terms of use, including those prohibiting use of web crawlers or spiders to gather data from the website, to determine that they related to use *of the website's dating services*. Defendant's use of the website to gather evidence for use in a patent lawsuit did not involve unauthorized uses of the dating services, and thus did not breach plaintiff's terms of use.[28]

Terms of use designed to prevent reproduction of website content also raise issues regarding whether such claims are preempted by copyright claims. Courts have generally declined to find claims for enforcement of such terms to be preempted by the Copyright Act, reasoning that terms of use restricting the manner by which a website can be accessed or used go beyond the protections provided under the Copyright Act. For example, in *Internet Archive v. Shell*, the Internet Archive sought dismissal on preemption grounds of the plaintiff's claim for breach of contract relating to Internet Archive's crawling and indexing of plaintiff's website in violation of terms of use that prohibited any copying of plaintiff's website for a ''commercial or financial purpose.'' The court rejected Internet Archive's preemption argument, finding that Internet Archive's alleged agreement to refrain from use of the material on plaintiff's website ''for commercial or financial purposes . . . lie[s] well beyond the protections [the website owner] receives through the Copyright Act''[29] (which, as discussed, allows for limited use of copyrighted content, *even for a commercial purpose*, if sufficiently transformative or unlikely to provide a substitute for the copyrighted work). The court reached this conclusion despite the fact that the Internet Archive is a non-profit entity – apparently on the basis of disputed allegations that Internet Archive's copying of the content at issue allowed it to ''acquir[e] . . . grant awards, donations, . . . and the expectation of acquiring additional intellectual property.''[30]

These cases suggest that future contractual disputes relating to web crawling or scraping for analytics purposes based on terms of use violations will likely focus on the proscriptions on automated data collection that are set forth in those terms of use.

3. Damages Relating to Unauthorized Data Collection.

The cases discussed above establish that website terms of use may be enforced against any party who accesses or uses a website in violation of those terms, and that, if sufficiently clear and unambiguous, those terms may prohibit any automated data collection from the website. However, a breach of contract claim also requires a showing of damages. To date, few of the cases involving breaches of contract relating to website terms of use have been decided on the merits. As a result, the issue of damages in such cases has received scant attention in reported case law. Those cases that have addressed the damages issue acknowledge the challenges and showing required to establish damages relating to violations of website terms of use.

Hines v. Overstock.com, Inc., 668 F. Supp. 2d 362, 367 (E.D.N.Y. 2009) (denying defendant's motion to stay or dismiss based on arbitration provision in its website's terms of service where defendant failed to rebut plaintiff's evidence that she was not aware of those terms, and that terms could only be seen by ''scrolling to the bottom of the screen-an action that was not required to effectuate her purchase'')).

[25] *See, e.g.*, *Cairo*, 2005 BL 9669 , at *2 (discussing website's terms of use prohibiting access to defendant's websites with ''any robot, spider or other automatic device or process to monitor or copy any portion'' of the websites).

[26] *Southwest Airlines*, 2007 BL 114340, at *1-2, *7.

[27] *Id.* at *8.

[28] *TrueBeginnings, LLC v. Spark Network Servs., Inc.*, 631 F. Supp. 2d 849, 853-56 (N.D. Tex. 2009).

[29] *Internet Archive v. Shell*, 505 F. Supp. 2d 755, 760, 767 (D. Colo. 2007).

[30] *Id.* at 769. *See also Craigslist, Inc. v. 3Taps Inc.*, No. CV 12–03816 CRB, 2013 BL 116811, at *11 (N.D. Ca. Apr. 30, 2013) (breach of contract claim against defendant that scraped user posts from Craigslist's website in violation of terms of use prohibiting automated access not preempted ''because the contract that Craigslist alleges here involves a number of extra elements not merely equivalent to rights under the Copyright Act'') (internal quotation marks omitted).

For example, in *Southwest Airlines*, the court granted summary judgment to Southwest on its breach of contract claim based on its finding that Southwest sufficiently demonstrated that defendant's services allowed Southwest customers to avoid the online check-in process, thereby decreasing web traffic to Southwest's website. By decreasing that traffic, the defendant deprived Southwest of valuable selling and advertising opportunities, and also interfered with Southwest's brand-building opportunities. Nonetheless, while Southwest established that it suffered some form of harm from the defendant's breach of the terms of use, the court declined to award any damages – finding that calculation of damages was ''impossible.'' Though it declined to award any damages, the court granted a permanent injunction in connection with Southwest' breach of contract claim.[31]

Indeed, because damages relating to violations of website terms of use may in some circumstances be difficult if not impossible to quantify, some courts have looked to liquidated damages provisions as an estimate of such damages. In *Myspace, Inc. v. The Globe.com*, MySpace alleged that the defendant used an automated script to send spam e-mails from various MySpace accounts established by defendant in violation of MySpace's terms of service providing that ''MySpace is for . . . personal use . . . only and may not be used in connection with any commercial endeavors,'' and which prohibited ''any automated use of the system'' or ''transmission of . . . spam[].'' MySpace's terms also provided that users agreed to pay $50 for each item of spam sent in violation of MySpace term's as ''a[n] . . . estimation of such harm.'' The court granted summary judgment on MySpace's motion for summary judgment on its breach of contract claim, and found that – because MySpace's actual damages from defendant's conduct was impracticable or extremely difficult to determine – liquidated damages of $50 per spam message was a reasonable measure of damages.[32]

The issue of damages is, of course, an intensely factual determination, but it should be noted that this issue is likely to play a key role in these cases in the future – with content owners trying to either quantify actual damages or establish the applicability of liquidated damages provisions, and those who use crawling and scraping tools arguing the impossibility of establishing such amounts. Based on the difficulty in establishing damages, content owners may also seek injunctive relief in such cases.

[31] *Southwest Airlines*, 2007 BL 114340, at *12.

[32] No. CV 06-3391-RGK(JCx), 2007 BL 64395, at *8-11 (C.D. Cal. Feb. 27, 2007). In so holding, the court noted that MySpace's costs relating to unsolicited spam messages included ''infrastructure costs, such as additional bandwidth,'' ''monitoring costs,'' and numerous ''hidden costs'' associated with ''deterrence (legal fees, software, etc.), depletion of customer goodwill, and liability implications associated with the unlawfully advertised product.'' *Id*. at *10. The court further relied on the fact that the federal statute prohibiting transmission of spam, the CAN-SPAM Act, sets statutory damages for unsolicited commercial emails at $25-300. *Id*.; *see also Craigslist, Inc. v. Naturemarket, Inc*., 694 F. Supp. 2d 1039, 1064-65 (N.D. Cal. 2010) (in connection with default judgment against defendant that sold software that enabled automated posting of ads on craigslist.com in violation of that website's terms of use, awarding liquidated damages of $100 per auto-posted ad as compensation for defendant's breach of contract).

C. Computer Fraud and Abuse Act.

Courts have also considered whether web crawling or scraping in breach of a website's terms of service constitutes a violation of the Computer Fraud and Abuse Act (''CFAA''), which prohibits access to a computer, website, server or database either ''without authorization'' or in way that ''exceeds authorized access'' of the computer.[33] While these terms have been variously defined, in essence, a person who accesses a computer ''without authorization'' does so without any permission at all, while a person ''exceeds authorized access'' where she ''has permission to access the computer, but accesses information on the computer that the person is not entitled to access.''[34] So long as a computer is publicly accessible, and not protected by password or other security measures, courts have declined to find any access of the website to be ''without authorization.''[35] Conversely, a CFAA claim may lie where a computer or website is protected from unauthorized access, either by technical measures or even explicit warnings in a cease-and-desist letter.[36]

Courts are split, however, as to whether access of a website in a manner prohibited by its terms of use ''exceeds authorized access'' of the website in violation of the CFAA. For example, in an early case on this topic, a federal court in Virginia granted summary judgment on AOL's CFAA claim based on the defendant's admission that it harvested email addresses from AOL's website in violation of its terms of use.[37] Several years later, in 2003, the Court of Appeals for the First Circuit seem-

[33] For example, the CFAA establishes criminal liability for whoever (1) ''intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains . . . information from a protected computer,'' 18 U.S.C. § 1030(a)(2)(C); (2) ''intentionally accesses a protected computer without authorization, and as a result of such conduct, recklessly causes damage,'' *id*. at § 1030(a)(4); and (3) ''intentionally accesses a protected computer without authorization, and as a result of such conduct, causes damage and loss,'' *id*. at § 1030(a)(5)(C). At least one state, California, has a ''functionally identical'' statute that likewise prohibits this conduct. *See e.g., 3Taps Inc.*, 2013 BL 116811, at *3-4 (describing California's Comprehensive Computer Data Access and Fraud Act, Cal. Penal Code § 502(e), as ''functionally identical'' to the CFAA, and holding that Craigslist stated claims for violation of both statutes by alleging that defendant accessed and scraped Craigslist's website despite cease-and-desist demands and technical measures deployed to prevent such access/scraping).

[34] *CollegeSource, Inc. v. AcademyOne, Inc.*, No. 10–3542, 2012 BL 279180, at *13-14 (E.D. Pa. Oct. 25, 2012) (quoting *LVRC Holdings LLC v. Brekka*, 581 F.3d 1127, 1133 (9th Cir. 2009)).

[35] *CollegeSource*, 2012 BL 279180, at *14 (finding that defendant's access of plaintiff's website for the purpose of scraping its course catalog was not ''without authorization'' because that website ''is available on the Internet and does not require a password or individualized access''); *Cvent, Inc. v. EventBrite, Inc.*, 739 F. Supp. 2d 927, 932-34 (E.D. Va. 2010) (defendant's scraping of content from plaintiff's website was not ''without authorization'' where website was publicly available and ''not protected in any meaningful fashion'').

[36] *See 3Taps Inc.*, 2013 BL 116811, at *4 (finding that Craigslist stated a claim against defendants that scraped its website after receiving a cease-and-desist letter prohibiting any access or use of Craigslist's website).

[37] *Am. Online, Inc. v. LCGM, Inc*., 46 F. Supp. 2d 444, 451 (E.D. Va. 1998) (''Defendants have admitted to utilizing software to collect AOL members' addresses. These actions were unauthorized because they violated AOL's Terms of Service.'').

   ISSN 1098-5190

ingly agreed with this theory by stating in *dicta* that ''[a] lack of authorization could be established by an explicit statement on a website restricting access.''[38]

These decisions, however, have been greeted with skepticism by later courts and commentators.[39] For example, in 2012, the Ninth Circuit, held in an *en banc* decision captioned *U.S. v. Nosal* that ''the phrase 'exceeds authorized access' in the CFAA does not extend to violations of use restrictions,'' but rather concerns ''hacking—the circumvention of technological access barriers.''[40] In reaching this decision, the Ninth Circuit emphasized the legislative history of the CFAA, noting that it was enacted in 1984 ''primarily to address the growing problem of computer hacking.''[41] The court further discussed the absurd results that would follow from criminalizing violations of website terms of use – *e.g.*, on dating websites that purport to require honest self-descriptions, describing ''yourself as 'tall, dark and handsome,' when you're actually short and homely, will earn you a handsome orange jumpsuit'' – and moreover, would allow for ever-shifting grounds for criminal liability as website terms of use are subject to change at any time, in any way, at the website owner's complete discretion. Thus, ''behavior that wasn't criminal yesterday can become criminal today without an act of Congress, and without any notice whatsoever.''[42]

While the current trend appears to be to reject broad theories that allow terms of use violations to be used as a basis to establish criminal liability under the CFAA (or analogous state statutes), this is a still an unresolved area in most circuits – and one that will likely further be argued in crawling and scraping cases.

D. Hot News Misappropriation.

In addition to asserting copyright claims based on incidental reproduction of copyrighted webpage material, numerous plaintiffs have asserted claims for hot news misappropriation relating to scraping of purely factual information. ''Hot news'' misappropriation – once a claim that existed under the federal common law, but which now exists only under the laws of five states[43] – provides a cause of action where a party reproduces factual, time-sensitive information that was gathered at the effort and expense of another party, and thereby deprives the gathering party of the commercial value of that information. Thus, for example, in *Int'l News Serv. v. Associated Press*, the Supreme Court in 1918 recognized a claim under federal common law for hot news misappropriation in connection with a wire service's republication of breaking news gathered by the Associated Press, which thereby deprived the Associated Press of the news value of its reporting.[44] The court justified its decision as protecting the ''quasi-property'' rights of profit seeking entrepreneurs who gathered time-sensitive information from those who would free-ride on the efforts of those entrepreneurs.[45]

Since hot news misappropriation generally concerns factual information rather than content that is subject to copyright protection, it is generally found not to be preempted by the Copyright Act.[46] However, courts have recognized hot news misappropriation as an extremely narrow claim that survives preemption only in very narrow circumstances that mirror the circumstances in *Int'l News Serv*. For example, in *Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, financial services firms alleged claims for copyright infringement and hot news misappropriation against a news aggregation website that reported on investment recommendations issued by the firms to their clients who paid to receive those recommendations before they became generally known to the investment community. On appeal from a denial of the defendant's motion to dismiss the hot news claim, the court found that plaintiff's claim was preempted by the Copyright Act. In so finding, the court emphasized that the plaintiffs' claim lacked an ''indispensable element of an *INS* 'hot news' claim,'' *i.e.*, ''free-riding by a defendant on a plaintiff's product, enabling the defendant to produce a directly competitive product for less money because it has lower costs.''[47] Rather, though the defendant's conduct potentially threatened plaintiffs' businesses, the defendant was actually *breaking* news generated by the plaintiffs' recommendations (and attributing the recommenda-

[38] *EF Cultural Travel BV v. Zefer Corp.*, 318 F.3d 58, 62-63 (1st Cir. 2003). In *EF Cultural Travel*, the First Circuit disagreed with the district court's finding that plaintiff's travel website was likely to prevail on its claim that its competitor violated the CFAA by scraping vacation pricing from its website; insofar as plaintiff's website was neither password protected nor governed by terms of use prohibiting scraping, defendants' conduct was not in exceed of the authorized access.

[39] *See, e.g.*, *Southwest Airlines*, 2007 BL 114340, at *14 (stating that ''[a] number of courts . . . have indicated . . . that a computer use which violates the terms of a contract made between a user and the computer owner . . . 'exceeds authorized access,' and hence violates the CFAA,'' noting that ''[t]hese cases, however, have received their share of criticism from commentators,'' and suggesting in dicta that defendant's access of Southwest's website did not exceed authorized access even though it violated Southwest's terms of use).

[40] 676 F.3d 854, 863 (9th Cir. 2012).

[41] *Id*. at 858.

[42] *Id*. at 861-62. The Fourth Circuit likewise declined, shortly after *Nosal* was decided, to find a CFAA violation where a computer system was accessed with authorization, even where the access was for an unauthorized purpose. *WEC Carolina Energy Solutions LLC v. Miller*, 687 F.3d 199, 206-07 (4th Cir. 2012) (employee who accessed his employer's network to misappropriate trade secrets in violation of company policy did not violate the CFAA where he was authorized to access the network).

[43] *See Pottstown Daily News Publ'g Co. v. Pottstown Broad. Co.*, 192 A.2d 657 (Pa. 1963) (Pennsylvania); *McKevitt v. Pallasch*, 339 F.3d 530 (7th Cir. 2003) (Illinois); *Nat'l Basketball Ass'n v. Motorola*, 105 F.3d 841 (2d Cir. 1997) (New York); *Pollstar v. Gigmania Ltd.*, 170 F. Supp. 2d 974 (E.D. Cal. 2000) (California); *Fred Wehrenberg Circuit of Theatres, Inc. v. Moviefone, Inc.*, 73 F. Supp. 2d 1044 (E.D. Mo. 1999) (Missouri).

[44] 248 U.S. 215, 245 (1918).

[45] *Id*. at 236; *see also Nat'l Basketball Ass'n*, 105 F.3d at 853 (''[*Int'l News Serv*.] is . . . about the protection of property rights in time-sensitive information so that the information will be made available to the public by profit seeking entrepreneurs. If services like AP were not assured of property rights in the news they pay to collect, they would cease to collect it. The ability of their competitors to appropriate their product at only nominal cost and thereby to disseminate a competing product at a lower price would destroy the incentive to collect news in the first place. The newspaper-reading public would suffer because no one would have an incentive to collect 'hot news.' '').

[46] *See Nat'l Basketball Ass'n*, 105 F.3d at 850-51 (''Courts are generally agreed that some form of [''hot news'' misappropriation] survives preemption.'') (citing *Fin. Info., Inc. v. Moody's Investors Serv., Inc.*, 808 F.2d 204, 208 (2d Cir. 1986)).

[47] 650 F.3d 876, 885-86, 901-02 (2d Cir. 2011).

tions to plaintiffs), rather than merely *repackaging* news that had been reported by plaintiffs.[48]

The *Barclays* case suggests the difficulty of stating a valid hot news misappropriation claim against a party engaged in automated data collection for purposes of data analytics. In many factual scenarios, scraping of information would not appear to qualify as ''free-riding'' within the meaning of *INS* so long as the scraper did not attempt to pass the information off as his own without attribution to the content provider. Indeed, many factual circumstances would appear similar to the recommendations at issue in *Barclays*, where the information is only valuable *because* it was attributed to the source. The fact that data analytics often involves the use of information to create entirely new insights (including in combination with information from other sources) suggests further difficulties in establishing the requisite ''free-riding,'' which under *Barclays* involves demonstrating that the underlying information was used to produce a directly competitive product.

E. Trespass to Chattels.

Courts have also recognized, in certain narrow circumstances, that unauthorized use of web crawling or scraping tools can give rise to a trespass to chattels claim, which ''lies where an intentional interference with the possession of personal property has proximately cause injury.''[49] For example, in *eBay, Inc. v. Bidder's Edge, Inc.*, eBay brought a trespass to chattels claim against the defendant, an online auction aggregation service that scraped auction information from eBay's website using spiders that accessed the website approximately 100,000 times per day in violation of eBay's terms of service and in defiance of cease-and-desist demands from eBay. eBay also moved to preliminary enjoin the defendant from accessing its website. In granting that motion, and finding that eBay was likely to prevail on its trespass to chattels claim, the court relied on the fact that defendant's spiders consumed a portion – albeit very small – of eBay's server and server capacity, and thereby ''deprived eBay of the ability to use that portion of its personal property for its own purposes.''[50]

In contrast, where tangible interference is absent, or is no more than theoretical or *de minimus*, courts have declined to recognize claims for trespass to chattel relating to the use of web crawling or scraping tools. For example, in *Tickets.com*, the court granted summary judgment dismissing Ticketmaster's trespass to chattel because Ticketmaster failed to present any evidence that its competitor's scraping of its website either caused physical harm to Ticketmaster's servers or otherwise impeded Ticketmaster's use or utility of its servers. In so holding, the court criticized the decision of the *eBay* court, and required a showing of ''some tangible interference with the use or operation of the computer being invaded by the spider.''[51] Later courts have generally agreed with the holding in *Tickets.com*.[52]

To the extent that *Tickets.com* presents the prevailing statement of law, and evidence of a tangible interference with a computer or server is necessary to state a claim for trespass to chattels based on unauthorized web crawling or scraping, courts are likely in the future to focus on evidence of tangible interference with systems.[53]

Conclusion

As indicated above, the legal landscape relating to web crawling and scraping is still taking shape—particularly insofar as few courts have considered claims based on crawling or scraping for analytics purposes. Further, because most cases involving the use of web crawling and scraping tools in other contexts have been highly fact specific, it is difficult to identify bright line rules for determining when use of such tools for analytics purposes is likely to give rise to liability. Nonetheless, the cases discussed above suggest a number of issues that should be considered both by website owners and by those who seek to perform analytics using data gathered from web-based sources.

These issues include (1) the language of the terms of use or service, and whether such terms address *access* to the website through automated means, *use* of any data collected through such means, and *use* of the website for anything other than the user's personal, noncommercial use; (2) the enforceability of the terms of use, for example, whether they are presented to the user through a clickwrap mechanism that requires the user to indicate his or her assent to those terms as opposed to a browsewrap agreement, or on a terms of use page that can be reached through a conspicuous link on every other page on the website and which indicates that any use of the website is subject to the user's agreement to those terms; (3) use of technological tools to deter unwanted crawling or scraping, including but not limited to the robots.txt protocols; (4) whether the website owner will license or authorize uses of content; (5) whether access to the website is protected such that a claim the CFAA or California's Penal Section 502 may be alleged; and (6) the extent to which the website content is protected by copyrighted.

Ultimately, while the claims and theories that may be advanced in connection with the use of web crawling

---

[48] *Id.* at 902-04. *Compare Barclays Capital*, 650 F.3d at 901-02, *with Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454, 460-61 (S.D.N.Y. 2009) (finding that the Associated Press adequately stated a claim for hot news misappropriation against defendant who rewrote news articles published by the Associated Press and passed them off as articles reported by defendant).

[49] *eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058, 1069 (N.D. Cal. 2000).

[50] *Id.* at 1062, 1071. *See also 3Taps Inc.*, 2013 BL 116811, at *14 (finding that Craigslist adequately stated a claim for trespass to chattels in connection with unauthorized scraping of its website where the defendant allegedly made ''mass copies tens of millions of postings from craigslist in 'real time' '').

[51] *Ticketmaster Corp. v. Tickets.com, Inc.*, No. CV997654HLHVBKX, 2003 BL 2425, at *5 (C.D. Cal. Mar. 7, 2003).

[52] *See, e.g.*, *Snap-on Bus. Solutions Inc. v. O'Neil & Assoc., Inc.*, 708 F. Supp. 2d 669, 679-80 (N.D. Ohio 2010) (denying the defendant's motion for summary judgment on plaintiff's trespass to chattels claim, finding that a reasonable trier of fact could find that defendant's scraping of plaintiff's website constituted trespass insofar as plaintiff put forth evidence that defendant's scraping tangibly interfered with plaintiff's use of its servers by causing those servers to crash).

[53] *See id.* at 681-82 (evidence that defendant's unauthorized scraping caused ''enormous spikes in . . . traffic'' to plaintiff's website, which spikes caused plaintiff's website to crash, required denial of defendant's motion for summary judgment on plaintiff's trespass to chattels claim).

   ISSN 1098-5190

and scraping tools for analytics purposes have yet to be deeply explored by courts, this is likely a temporary state of affairs. Rather, given the increasing number and availability of tools for aggregation and analysis of content in the Big Data era, courts will ultimately be required to address these complicated issues.