Hedge Fund Law Report
An Acuris company

Intellectual Property

Best Practices for Private Fund Advisers to Manage the Risks of Big Data and Web Scraping

Jun. 15, 2017

By Robert G. Leonard, Jeffrey D. Neuburger and Joshua M. Newville, *Proskauer Rose LLP*

On April 13, 2017, craigslist obtained a judgment against RadPad, a third party that collected data through automated means from its site. The $60.5 million judgment was based on various claims relating to RadPad's use of sophisticated techniques to evade detection and harvest content from craigslist's site, as well as distribution of unsolicited commercial emails to craigslist users to market RadPad's own apartment rental listing service.[1] While it is doubtful that craigslist will ever collect this sizeable judgment, the case highlights some of the issues faced by persons, such as hedge fund managers, who collect – or engage others to collect – data through automated means for commercial purposes.

This article provides an overview of big data and web scraping, outlines potential sources of liability to hedge fund managers that collect big data and describes best practices for navigating several areas of potential liability.

Introduction to Big Data and Web Scraping

Automated data collection – also referred to as web scraping, data scraping and spidering, among other names – is a practice that has been controversial since the early days of the commercialization of the internet. The practice refers to the use of a "robot" (or "bot" or "spider") to collect data from websites. Bots may target specific websites on a periodic basis or crawl through the web more generally. Data collected through these means can be used for many different purposes, including by hedge fund managers to analyze prospective and existing investments.

Website owners typically take certain measures to inform third parties about their preferences regarding scraping website data. Most commonly, a site's terms of service will contain language stating the site's policy regarding scraping; often, the terms generally prohibit automated database scraping activities. Also, a website's underlying code will generally communicate with search engine spiders and data scraping bots by placing a "robots.txt" file on its site that may signal that certain pages are off-limits to automated scraping, or a crawl delay to limit how many times a spider can access the site per minute.

When apprised of certain unwanted data-collection activities, website owners might institute certain technical measures, including IP address blocks, and might expressly revoke a third

party's website access by sending a "cease-and-desist" letter to the entity behind unwanted scraping.

Disputes may arise when website owners seek to prevent third parties from, among other things:

1. extracting content from the relevant site for the third parties' commercial or competitive use;
2. copying content protected by copyright;
3. extracting content in contravention of the site's terms of service or technical measures; or
4. disrupting the site's operations through scraping activities (*e.g.*, causing a website outage through crawling excessively, spamming end-users or causing the site owner to incur unwanted IT-related costs).

Despite the relative maturity of e-commerce, the legality of automated data collection is still unsettled. While there have been many cases that have examined scraping disputes under various state and federal statutes, the law is not uniform, and past decisions have been fact-specific in nature.

Sources of Liability

Breach of Contract

Most websites' terms of use, terms of service or end-user license agreements (EULAs) typically include language prohibiting automated data collection.[2] Violation of the EULA by, for example, scraping information has been used successfully as the basis for breach-of-contract claims.

In some instances, site users are required to register and expressly agree to the site's terms before being permitted to access the principal areas of the site. In many cases, however, the EULA or site terms are purportedly accepted by a user simply by accessing the site. This type of electronic contracting is often referred to as a "browsewrap" agreement (*e.g.*, a link to the terms of service is included at the bottom or top of each web page but without any "click to accept" or "clickwrap" feature that is typical of many e-commerce transactions and social media registrations). While courts are often skeptical about browsewrap agreements in consumer transactions, they evaluate the validity of these agreements based on whether the user had actual or constructive knowledge of a website's terms and conditions and whether the user manifested assent to those terms.

Whether the site terms are enforceable against a sophisticated commercial party (as opposed to a consumer) may also be relevant. Depending on the presentation, one may be able to argue that a particular website's terms are unenforceable, although as the U.S. Court of Appeals for the Second Circuit once ruled, such a contention can falter when a commercial entity that may not have been aware of the website terms the first time it scraped data, was found to have been aware of the terms after multiple contacts with the site.[3] The argument that one had no notice of the terms is naturally harder to make with respect to clickwrap or click to accept forms of EULAs, which are more consistently found to be enforceable.

Computer Fraud and Abuse Act

Upon noticing unwanted scraping, a site owner may take technical and legal steps to discourage that activity. Data collectors who ignore these impediments and continue to scrape risk running

afoul of certain federal laws. The Computer Fraud and Abuse Act (CFAA), while generally a criminal computer hacking statute, prohibits access to information from a computer, website, server or database that is "without authorization" or in a way that "exceeds authorized access." The CFAA is nearly always asserted by site owners as the basis for relief against data collectors.

To bring a successful claim under the CFAA in a data-scraping case, site owners must, among other things, advance evidence that shows the scraper intentionally accessed the site "without authorization." Considering that most websites are generally open to the public, a site owner will be required to show that the site revoked the third-party data scraper's permission to access the site in question and that the scraper nevertheless continued to access it. Beyond terms of service that prohibit scraping, revocation or "de-authorization" typically takes the form of a cease-and-desist letter prohibiting further site access, along with technological impediments such as IP address blocks against a particular data scraper.

In recent years, courts have held that a site owner has communicated a complete revocation of access when it sends a cease-and-desist letter that forbids website access for any reason and then backs up that action with a technological barrier. Indeed, following recent court rulings, site owners are not merely relying on the enforceability of their site terms alone, but instead expressly revoking access to unwanted data scrapers. The U.S. Court of Appeals for the Ninth Circuit recently bolstered the importance of express revocation when it ruled that "a violation of the terms of use of a website – without more – cannot be the basis for liability under the CFAA."[4]

Securities Laws

The use of automated data collection for investment-research purposes may also give rise to issues under securities laws. First, hedge fund managers that obtain or receive data collected as a result of web scraping may come into possession of material nonpublic information (MNPI). If the information has not been disseminated broadly to the market and would be considered important by a reasonable investor in making an investment decision, it may be considered MNPI.

Second, if the data were collected in a manner considered "deceptive," then there is a risk that trading on that information may be considered part of a fraudulent scheme in violation of the anti-fraud provisions under the securities laws. Behavior in violation of an "ever-present duty not to mislead" may violate these provisions even when the trader is under no duty to the source of the information.[5] Circumventing security protocols; disguising or failing to reveal a scraper's identity on a site (where required); and simulating human transactions, among other behaviors, each could be viewed as an affirmative misrepresentation constituting a "deceptive device" under Section 10(b) of the Securities Exchange Act of 1934, which could form the basis for such a fraud claim.

Finally, even where data collectors fully comply with the terms of a site's agreements and its security protocols, state attorneys general may raise concerns under state laws about practices that take "unfair advantage" of access to information and practices that are against public policy generally.[6]

See "Supreme Court's Ruling in *Salman v. U.S.* Affirms the Importance of a Tipper's 'Personal Benefit' for Insider Trading, but Also Creates Uncertainty" (Feb. 9, 2017); "General Insider Trading Policies and Procedures May Be Insufficient for Hedge Fund Managers to Avert SEC Enforcement Action" (Nov. 3, 2016); and "Lessons for Hedge Fund Managers From the Government's Failed Prosecution of Alleged Insider Trading Under Wire and Securities Fraud Laws" (Jul. 21, 2016).

Copyright Infringement

In certain circumstances, automated data collection may infringe upon a site owner's copyright or other intellectual property rights. Under the Copyright Act, copyright protection is embedded within original works that are fixed in any tangible medium, which can certainly include website content, images, or in some cases, portions of the underlying website code. Therefore, because web-scraping tools generally index information on a targeted webpage, regardless of the type of information the tools are seeking to obtain, if that web scraping leads to the reproduction of any copyrighted content, the activity may give rise to a claim for copyright infringement.

With respect to websites containing user-generated content, a data scraper may be unaware that some site owners may have previously obtained an exclusive license to, or actual assignment of copyright ownership in, the content posted by users. In those circumstances, the website operator would then have standing to bring a copyright claim against entities scraping and copying that content.[7] However, compelling users to grant the site owner an exclusive license in the EULA is not a widespread approach, as sites generally do not garner exclusive copyright rights in user content for both practical and business reasons.

See "How Hedge Funds Can Protect Their Brands and IP: Pepper Hamilton Attorneys Discuss Trademarks and Copyrights (Part One of Two)" (Feb. 23, 2017).

Additional Issues

Hedge fund managers should be aware of certain additional issues that may arise from using automated data collection:

- Excessive automated data collection can interfere with the performance of a site. To the extent a site crashes, an end user experiences delays or a site's operational capacity is otherwise burdened, the data collector may be deemed to have interfered with the site owner's use of its tangible property. In these circumstances, site owners have brought trespass to chattels claims. Technological protocols such as robots.txt instructions regarding frequency and depth of scraping are intended to mitigate the risk of interfering with a site's operational capacity.
- Similar to liability under the CFAA, circumvention of technological control measures, such as measures to block automated access – such as a completely automated public Turing test to tell computers and humans apart (CAPTCHA) and "I am not a robot" measure – can also create the basis for liability under the Digital Millennium Copyright Act of 1998 (DMCA). The DMCA provides that "no person shall circumvent a technological measure that effectively controls access to a work protected under this title."[8]
- The Consumer Financial Protection Bureau (CFPB) recently called for comments on access to financial data via web scraping by FinTech companies, including data providers.[9] In its comment letter, the American Bankers Association (ABA) recommended that the CFPB ensure that data aggregators be held to the same data protection and notification standards as banks. Specifically, the ABA recommended that consumer data be subject to the protections of the Gramm-Leach-Bliley Act, which affords certain protections to "non-public personal information." [10]
- Collected data may contain personally identifiable information (PII). Generally, PII is data that is considered personal in nature and could be used to compromise the privacy of an individual. PII includes sensitive and nonpublic financial, health or other data or attributes,

such as addresses or financial account numbers. Data collectors should take care to anonymize data that comes from servers that contain PII. Even hedge fund managers who purchase scrubbed data from third parties should check to ensure the information they receive is fully anonymized and, if not, take steps to remove all identifying information. See "SEC Enforcement Action Illustrates Focus on Investment Adviser Obligation to Secure Client Information" (Jun. 23, 2016).

Hedge Fund Manager Best Practices

To minimize the risk of encountering any legal and compliance pitfalls, hedge fund managers can follow certain best practices in collecting big data as part of their investment research, either directly or through the use of a third-party data provider. Best practices include:

1. *Abide by the website terms.* Although, in practice, it is impractical for a data collector to review the terms of every website that it accesses, it is important to remember that there can be some risk to engaging in such activities without conducting such a review, especially to the extent that particular terms of service prohibit automated data collection on a site.
2. *Follow technological protocols, and refrain from circumventing impediments to automated data collection.* Some sites use, for example, the robots.txt protocol to specify (1) which bots are allowed to access the site, if any; (2) which bots are blocked from the site; and (3) pages of the site that are available for access, including frequency of access. Compliance with the robots.txt protocol is voluntary, although customary in the industry, and has previously been considered by courts when considering a site's stance toward web crawlers. Other sites use impediments such as blocking the IP address of a specific data collector or other mechanisms to authenticate human users. Common techniques include CAPTCHA and image recognition tests.
3. *Collect factual information; avoid expressive and proprietary content.* Bots used by hedge fund managers should have the ability to avoid collecting any information with creative elements, including, in particular, images. They should also avoid manipulating creative elements of the site in the course of collecting any factual information. Keep in mind that, beyond what is expressly protected content, the site owner may consider other information to be proprietary and confidential, particularly in the context of a bot collecting that information as part of a larger data set.
4. *Avoid collecting PII.* Anonymize data that comes from servers that contain PII. This includes making sure that data provided by a third party has also been appropriately scrubbed.
5. *Evaluate any potential or existing relationships with third-party data providers.* Factors to consider in determining whether to engage the services of a third-party data provider include protection from legal claims, downstream communication of intellectual property and any potential loss of control over compliance practices. Additionally, it is important to conduct careful diligence of the vendor, including inquiring as to the source of the data and whether it has been appropriately anonymized; review the contract with the vendor and assess any potential areas of liability; and monitor the vendor's activities on an ongoing basis. Additional consideration should be given when a data collection project by a vendor is the type of activity it regularly performs for any customer or whether it is a "custom" job; this consideration may implicate whether a vendor is a company's agent and affect considerations

of contributory or vicarious liability should a vendor be found liable for its data collection practices.

6. *Leave a paper trail.* Whether reviewing the policies and procedures for in-house analysts or conducting formal due diligence of data aggregators who provide customized or non-exclusive services, compliance officers should have policies and procedures in place to monitor key risks associated with automated data collection.
7. *Stay informed.* In addition to automated data collection, hedge fund managers should be aware of a wider set of data-science issues, including issues arising from, among others, use of satellite imagery, use of drones, location tracking from cell phones and privacy issues generally.

Robert G. Leonard is a partner in Proskauer's hedge funds group. For more than 25 years, Leonard has been structuring, organizing and representing hedge funds, funds of funds and other private investment funds (both domestic and offshore) and investment advisers.

Jeffrey D. Neuburger is a partner at Proskauer, co-head of the technology, media & telecommunications group, a member of the privacy & cybersecurity group and editor of the firm's new media and technology law blog.

Joshua M. Newville is a partner in Proskauer's litigation department and a member of the private equity and hedge fund litigation team. Newville handles securities litigation, enforcement and regulatory matters, representing corporations and senior executives in civil and criminal investigations. Previously, he served as senior counsel in the Asset Management Unit of the SEC's Division of Enforcement.

---

[1] *See Craigslist, Inc. v. RadPad, Inc.*, No. 16-01856 (N.D. Cal. Apr. 13, 2017). In this case, which was a default judgment, craigslist revived a dispute against an aggregator (3Taps) by filing a complaint against RadPad, a real estate listing site that had allegedly received scraped craigslist data from 3Taps before the *3Taps* case was settled. In its complaint, craigslist claimed that after the *3Taps* litigation was settled, RadPad and its agents began their own independent efforts to scrape craigslist's site, despite receiving a cease and desist letter from craigslist barring RadPad from using the craigslist site. Craigslist alleged that RadPad scraped thousands of user postings and thereafter harvested users' contact information to send spam in an effort to entice users to switch to RadPad's services. *See Craigslist, Inc. v. RadPad, Inc.*, No. 16-1856 (N.D. Cal. filed Apr. 8, 2016). Craigslist obtained relief under several causes of action, including copyright infringement, breach of contract, CAN-SPAM and the Computer Fraud and Abuse Act and the California state law equivalent.

[2] For example, craigslist's EULA prohibits, among other things, the use of "[r]obots, spiders, scripts, scrapers, crawlers" and the transmission of any "misleading, unsolicited, unlawful, and/or spam postings/email."

[3] *See Register.Com, Inc. v. Verio, Inc.*, 356 F.3d 393 (2nd Cir. 2004).

[4] *See Facebook, Inc. v. Power Ventures, Inc.*, 828 F.3d 1068 (9th Cir. 2016).

[5] *See SEC v. Dorozhko*, 574 F. 3d 42 (2nd Cir. 2009).

[6] See "Thomson Reuters to Suspend Early Data Survey Release" (Jul. 8, 2013), available at http://www.reuters.com/article/us-thomsonreuters-consumerdata-idUSBRE9670WE20130708.

[7] See generally *Metropolitan Regional Info. Sys., Inc. v. American Home Realty Network, Inc.*, 722 F.3d 591 (4th Cir. 2013).

[8] See, *e.g.*, *Ticketmaster L.L.C. v. RMG Technologies, Inc.* 507 F. Supp. 2d 1096 (C.D. Cal. 2007). In this case, a company marketed web-based software tools that enabled ticket brokers to repetitively access the Ticketmaster website and acquire large numbers of tickets very quickly. Ticketmaster successfully alleged a violation by the company of the anticircumvention provisions of the DMCA for bypassing Ticketmaster's CAPTCHA technology.

[9] *See* "Request for Information Regarding Consumer Access to Financial Records" (Nov. 14, 2016).

[10] *See* "ABA Makes Recommendations to Protect Customer Data When Shared" (Feb. 21, 2017).

This material has been printed by and is for their consumption only. The full Terms of Use are available at www.hflawreport.com.
UNAUTHORIZED USE OR DISTRIBUTION IS UNLAWFUL