Portfolio Media. Inc. | 111 West 19th Street, 5th Floor | New York, NY 10011 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

The Latest Legal Trends In Data Scraping And Ownership

By **Wesley Horner and Bart Eppenauer** (August 17, 2020, 1:51 PM EDT)

Wesley Horner

Bart Eppenauer

For many people, the best place to start learning is the internet. From casual browsing to competitive intelligence, publicly available data is one of our best sources of information. With the explosion of artificial intelligence, scraping data from public websites to train AI models has become the new normal. But to some, data scraping almost seems like hacking.

In May, the U.S. Court of Appeals for the Eleventh Circuit ruled in Compulife Software v. Newman that scraping publicly available data can be an "improper means" for acquiring a trade secret, and the U.S. Supreme Court is considering a petition for a writ of certiorari, in LinkedIn Corp. v. HiQ Labs Inc., which addresses whether scraping publicly available data can violate the federal anti-hacking law.

As competitors and courts grapple with different fact patterns, data owners are finding that the law does not provide clear guiding principles for what kind of scraping is legal and what kind is off-limits. So what should you do if you find out someone is scraping your data, or if you want to do the scraping? It really depends on the type of scraping.

What exactly is data scraping?

Data scraping — sometimes called text and data mining — means extracting information from a website, database or program. It could be done manually or using automated software.

For example, web scraping can be performed with software that extracts information displayed on webpages, listens to data feeds from servers, or even performs higher level analytics like computer vision and natural language processing to identify relevant web content for later analysis.

Although data owners knowingly publish web content, large-scale scraping often occurs without the knowledge or consent of the data owner. The extracted information may be publicly available, licensed, available through some type of account, or accessed in some other way.

Why do people scrape data?

Data scraping is often done to obtain useful data or analytics. For example, data scraping is one way to

gather information for a dataset, which can be mined for insights. The EU Copyright Directive defines text and data mining as "any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[1] Someone might consider using data scraping to obtain some kind of commercial advantage, to provide some kind of service, for research purposes, or even with malicious intent.

Some types of scraping are commercially or socially accepted. For example, Google crawls[2] public webpages to support its search engine, which can increase traffic to the scraped website. Other types of scraping such as price monitoring,[3] sentiment monitoring,[4] aggregation[5] and predictive[6] analysis are prevalent, but can lead to commercial disputes, often when the scraper is a competitor.

Data scraping can also be done to support predictive analysis for research efforts, such as training a machine learning model.[7] Advancements in AI are increasingly limited by the lack of relevant training data. There is also a growing open-source community that values open sharing of information. These two influences can lead researchers to scrape web content, develop datasets and publish results.

Depending on the application, different types of data might be of interest. Some might be interested in gaining insights from creative content like articles, commentary, images or videos, proprietary information like prices or insights, personal data like profiles or contact information, or other content displayed or available from webpages.

Datasets and models are becoming increasingly valuable, and many researchers and commercial entities make their datasets and models publicly available, subject to some governing license (which may be an open source license). Companies like Microsoft Corp. have begun focusing on removing barriers to data innovation and open data sharing.[8]

What do you do if someone is scraping your data?

There are a number of reasons a data owner may not want its data to be scraped, from maintaining a competitive advantage to protecting its data or creative content. One way data owners can try to protect their data is through technical barriers. The reality, however, seems to be that most technical barriers can be circumvented. The best way to prevent others from using your data is simply not to publish it, although in many cases this is not a realistic option.

There are some potential legal barriers to data scraping, so a data owner might investigate possible causes of action to try to stop someone from scraping data. However, not all of causes of action will apply, depending on the nature of data being scraped, the relationship with the data scraper, the type of harm caused, and the different laws and jurisdictions. Compounding the challenge, the law is evolving, so it is not always clear whether data scraping actually violates the law.

Computer Fraud and Abuse Act

Take the Computer Fraud and Abuse Act, for example, which prohibits obtaining information by "intentionally access[ing] a computer without authorization or exceed[ing] authorized access."[9] Whether data scraping is done "without authorization" or "exceeds authorized access" has been the subject of a few recent court cases. One emerging distinction is between free, publicly available data on the one hand, and protected data on the other.

In its 2016 decision in Facebook Inc. v. Power Ventures Inc., the U.S. Court of Appeals for the Ninth

Circuit found a violation where a data aggregator accessed Facebook data using usernames and passwords provided by Facebook users to the aggregator, Facebook sent a cease-and-desist letter, but the aggregator continued without authorization from Facebook.[10]

By contrast, in the closely watched case hiQ Labs v. LinkedIn, the Ninth Circuit found that when a website grants the public access to its data without requiring any kind of authorization such as a password, scraping this type of public data most likely does not constitute access without authorization in violation of the CFAA.[11] LinkedIn recently petitioned the Supreme Court on this issue,[12] and the court will likely decide whether to hear the case in the next few months.

In the meantime, data owners could consider password-protecting their data, as there may be recourse under the CFAA when password-protected data gets scraped.

Copyright Infringement

Another possible cause of action for data owners is copyright infringement. When a webpage is accessed using a web browser, the browser caches the webpage. Some courts have found that this type of caching falls within the Copyright Act's definition of "copy."[13] The terms of service on a webpage most likely grant a license to cache the webpage, unless the manner of access violates the terms. So if a website prohibits data scraping in their terms of service, data scraping that uses caching might constitute copyright infringement.[14]

There may be also a copyright claim predicated on making a copy of the scraped data. However, the data needs to have sufficient originality and creativity, and copyrightable elements need to have been copied.

In a recent case, UAB "Planner5D" v. Facebook Inc., data owner Planner5D sued Princeton and Facebook over a collection of 3D models. Planner5D alleged that researchers at Princeton improperly scraped its collection of 3D models and compilations of household objects and room scenes. The district court dismissed Planner5D's initial copyright claims in part because Planner5D had not sufficiently alleged its models were made with the requisite level of creativity, as opposed to simply being exact renderings of real-world objects.[15]

Some uses of copyrighted works may constitute fair use. For example, in 2015 the U.S. Court of Appeals for the Second Circuit found in Authors Guild Inc. v. Google Inc. that Google's digitization of copyrighted works for the Google Books program was a noninfringing fair use.[16]

Some commentators have suggested that using copyrighted works to train an AI model may also constitute fair use.[17] However, under the current legal framework, the applicability of fair use has not been definitively decided.

Trade Secrets

Some data owners have argued their data is a trade secret, and scraping their data constitutes improper acquisition. This type of claim usually will not work when the scraped data was publicly available, since a trade secret has to actually be a secret. However, it may not be as easy to show lack of secrecy as it may seem.

In the Planner5D case, Planner5D also alleged that downloading its collection of 3D models constituted a

misappropriation of trade secrets. Planner5D's amended claims survived a recent motion to dismiss on the basis that the question of secrecy was a factual challenge to the allegation that reasonable measures were taken to protect the trade secret.[18] Even if it made its files publicly available, Planner5D alleged the use of both structural measures (secret internet addresses) and legal measures (prohibitions in the terms of service on accessing or acquiring the files) to protect its files, which was sufficient to survive a motion to dismiss.[19]

In a recent post-trial appeal, the Eleventh Circuit emphasized the fact that data is publicly available does not automatically resolve a trade secret claim. In Compulife Software v. Newman, the defendants hired someone to scrape 43 million insurance quotes from a competitor's public website.[20]

In a decision that reads almost like a spy novel, the court characterized this as a "scraping attack" by a "hacker," even though the insurance quotes were freely available to the public.[21]

> [T]he scraped quotes were not individually protectable trade secrets because each is readily available to the public — but that doesn't in and of itself resolve the question whether, in effect, the database as a whole was misappropriated. ... Nor does the fact that the defendants took the quotes from a publicly accessible site automatically mean that the taking was authorized or otherwise proper.[22] [23]

While an owner of a public website "has plainly given the world implicit permission to access as many quotes as is humanly possible, using a bot to collect an otherwise infeasible amount of data may well be [improper]."[24] With this guidance, the court remanded to the lower court for reconsideration.

In either of these cases, it is too soon to say how the courts will resolve the trade secret claims. However, the cases suggest data scrapers should be cautious in deciding whether and how much to scrape.

Trespass to Chattel

Another cause of action data owners have invoked is trespass to chattel. The underlying concept is that the scraping has somehow interfered with the data owner's possessory interest in its computer system. The major limitation with this cause of action is the requirement to show tangible harm. This type of claim is more likely to work where the number of incoming requests takes down a server or causes some other kind of demonstrable harm to a computing system.

Breach of Contract

Another possible liability theory is breach of contract based on a website's terms of use. For example, if the terms prohibit scraping or the use of automated tools, and the data scraper assented to the terms, scraping may constitute a breach of contract.

The major uncertainty with this cause of action is enforceability, which usually turns on whether there was actual or constructive notice and assent. Not only are there different rules in different states, in any particular state, there is often no clear answer for a given set of facts. Some of the factors that may impact enforceability include whether a website's terms of use are only made available through an optional link, whether the link is sufficiently conspicuous, whether there was some sort of express manifestation of assent (like a check box that references the terms), or even the level of sophistication of the data scraper.

Like trespass to chattel, a successful breach of contract claim requires the website owner to show some tangible harm to establish damages.

What do you do if you want to scrape data?

When evaluating whether data scraping is legal, consider the type of data to be extracted, in which jurisdiction the proposed activities will take place, the purpose of the proposed data scraping, and the manner in which data will be extracted.

Given the changing landscape and patchwork of laws across different jurisdictions, the only way to be truly sure data scraping is allowed may be to negotiate a license. Otherwise, data scraping may be legal when the terms of service for a website do not prohibit scraping, or an open-source contributor provides a blanket license for content being distributed. Data scraping could be in violation of some law when it is done without authorization, or when the volume or nature of the activity negatively impacts server functionality or availability.

However, there are a significant number of circumstances where data scraping is not clearly legal or illegal. In that situation, what can you use to guide your decision-making? One possible consideration is internal consistency. If you are in a position to potentially be both a data scraper and a data scrapee, you may want to make sure different branches of your organization are in accord. Another possible consideration is custom. How are similarly situated others acting? Of course, custom may not be dispositive, but it does tend to inform the law.

In light of the uncertainty in a growing number of situations, Congress could step in and provide guidance. Other countries have already taken legislative action in this arena.

For example, Article 3 of the EU Copyright Directive[25] provides an exception to copyright and database rights for text and data mining carried out for the purposes of scientific research. Article 4 of the EU Copyright Directive expands the exception to any entity, but specifically permits copyright owners to opt-out of the application of the exception. Considering even commercial uses of text and data mining can serve the public good,[26] with the potential for AI to solve some of our most pressing societal issues, Congress should consider providing clarity around text and data mining for machine learning, like in Japan[27] and Europe.

In much the same way that technology and big data have us rethinking some fundamental concepts like data privacy and cybersecurity, we need much greater legal clarity on the boundaries of property rights for data ownership and permissible uses of public data.

Wesley Horner is a senior associate at Shook Hardy & Bacon LLP.

Bart Eppenauer is a managing partner at the firm. He is the former chief patent counsel for Microsoft.

[1] EU Directive on Copyright and Related Rights in the Digital Single Market (2019), art. 2.1(2) ("EU

Copyright Directive").

[2] How Search organizes information, available at https://www.google.com/search/howsearchworks/crawling-indexing/.

[3] See Top 10 Price Monitoring tool in 2020, available at https://www.octoparse.com/blog/top-10-price-monitoring-tool.

[4] See Empowering Analysis with Online Public Sentiment Data, available at https://scrapinghub.com/extra-use-cases/public-sentiment-data/.

[5] See Williams, Janet, How Web Scraping for News Aggregation Works, available at https://www.promptcloud.com/blog/web-scraping-for-news-aggregation/.

[6] See Larson, John, The Predictive Power of Web Scraped Product Data for Institutional Investors: a GoPro Case Study, available at https://blog.scrapinghub.com/gopro-study.

[7] See Web Scraping for Training Data, available at https://www.promptcloud.com/web-scraping-for-training-data/.

[8] See Closing the Data Divide: The Need for Open Data, available at https://news.microsoft.com/opendata/.

[9] 18 U.S.C. § 1030(a)(2)(C).

[10] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067–68 (9th Cir. 2016), cert. denied, 138 S. Ct. 313 (2017).

[11] 938 F.3d 985, 1003 (9th Cir. 2019) ("It is likely that when a computer network generally permits public access to its data, a user's accessing that publicly available data will not constitute access without authorization under the CFAA."), petition for rehearing denied, 17-16783 (9th Cir. Nov. 8, 2019).

[12] LinkedIn Corp. v. HiQ Labs, Inc., Petition for a Writ of Certiorari, No. 17-16783 (Mar. 9. 2020) (presenting the question, "Whether a company that deploys anonymous computer 'bots' to circumvent technical barriers and harvest millions of individuals' personal data from computer servers that host public-facing websites—even after the computer servers' owner has expressly denied permission to access the data—'intentionally accesses a computer without authorization' in violation of the Computer Fraud and Abuse Act'").

[13] See, e.g., Ticketmaster v. RMG Techs., Inc., 507 F. Supp. 2d 1096, 1105 (C.D. Cal. 2007).

[14] Id. at 1106–1110.

[15] UAB "Planner5d" v. Facebook, Inc., No. 19-cv-03132-WHO, 2019 WL 6219223, at *8 (N.D. Cal. Nov. 21, 2019).

[16] The Authors Guild Inc., et al. v. Google, Inc., 804 F.3d 202, 229 (2d Cir. 2015).

[17] See, e.g., Robinson, Karen, Copyrights in the Era of AI, available

at https://theblog.adobe.com/copyrights-in-the-era-of-ai/.

[18] UAB "Planner5d" v. Facebook, Inc., No. 19-cv-03132-WHO, 2020 WL 426073, *6–7 (N.D. Cal. July 24, 2020).

[19] Id. at *7.

[20] Compulife Software v. Newman, 959 F.3d 1288, 1300 (11th Cir. 2020).

[21] Id. at 1299-1300, 1314.

[22] Id. at 1314.

[23] Id.

[24] Id.

[25] See EU Copyright Directive, available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790&rid=1.

[26] See AI for Good, available at https://www.microsoft.com/en-us/ai/ai-for-good.

[27] See Japan Amends its Copyright Legislation to Meet Future Demands in AI and Big Data, available at http://eare.eu/japan-amends-tdm-exception-copyright/.