CHAPTER 3

Reusability of Publicly Accessible User Data on Platform Websites

Haewon Chung

Abstract

The open data movement has been concerned with increasing access to public sector data. In the future of open data, governments should also consider the reuse of user-generated data on popular online services and third-party use of automated programs to extract publicly accessible data from these platforms. Internet users increasingly rely on popular platforms, such as Facebook, Twitter, and LinkedIn, to access information and to communicate with others. This emerging structure of the virtual world allows platform companies to occupy an advantageous position over third parties seeking access to user-generated data. Platforms can deploy various legal and technological barriers against third-party access. Third-party use of publicly accessible data on the Internet can spur various commercial and non-commercial developments. Legal intervention is needed against platforms' proprietary management of such data for profit-maximization because their practice impedes the Internet as an open and generative technology and deters progress in society. Data are a valuable resource in today's knowledge-driven society. Institutions committed to the open data approach must also improve the reusability of web data.

Acknowledgements

I want to thank Dr. Teresa Scassa for providing valuable comments on this chapter. This research was carried out as part of the Geothink partnership, and the support of the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

In the last decade, the global open data movement has largely focused on opening up third-party access to public sector data produced and collected by the government. Open government data policies and legislation increase government transparency (GoC, 2017). They also promote innovation, research, and competition by allowing others to access and use public sector data. Governments can also encourage economic growth and public benefit by improving third-party access to publicly accessible web data. Internet users upload and publicly share a massive amount of data and information through digital intermediaries, such as Facebook, Twitter, and Instagram (Constantine, 2012). However, platform companies' legal and technological access barriers discourage third-party data users. In the future of open data, governments should support the reuse of publicly shared data on online platforms with data policies and legislative measures that remove unnecessary barriers to third-party data use.

Third-party data users can access and gather data directly from platform websites manually or by using an automated program (i.e., a bot). This process is known as data scraping. Third parties can apply publicly accessible factual user data in various commercial and non-commercial endeavours, such as creating new products and services and conducting research on society and technology. On the other hand, platform companies, which often depend on advertising profits, can limit competition and maximize profit by tightly controlling user-generated content. Allowing platforms to turn publicly accessible user data into a private resource is not in the interest of the general public. Moreover, such tactics contradict the open nature of the Internet and its tremendous capacity to encourage innovation and knowledge.

Governments need to improve the regulation of web data. They should not leave it up to the oligopolistic market on the web to propertize valuable innovation resources such as web data. Relevant laws need to be modernized and legal uncertainties should be removed to

promote fair and transparent third-party use of publicly accessible web data. Since Internet users share a variety of content over the Internet, I will limit the discussion in this chapter to the use of publicly shared factual data (e.g., user profiles and locational data). For the remainder of this chapter, I will refer to publicly shared factual data hosted on platform websites as "public user data." I will use the term "user-generated content" to refer to broader user contributions online that include factual data and copyright-protected works (e.g., original written expressions and images). It should be noted that not all user-generated content online is publicly accessible. Some platforms allow users to privately share data and information with one or more users, and such content is not publicly accessible. This chapter does not consider the use of private data.

This chapter is intended to encourage discussions among Internet users, scholars, and lawmakers about the Internet as an open network, automated data scraping, and web data regulation that can promote new technology and public benefit. The remainder of this chapter is organized as follows. Section one describes the role of platform services, possible uses of public user data, applicable laws on data access and use, and the platform businesses' possible motivation in user data regulation. Section two examines how new innovation policy in law supported the emergence of the Internet as an open network and a generative technology. Lawmakers need to take an ongoing and active role in protecting the Internet's open design as well as data and information freedom on the Internet. Following this, section three examines data scraping and anti-bot technologies. It then reviews contract law, tort law, copyright, and anti-circumvention law, which platform companies may rely on to prevent third-party data scraping. Section four discusses why it is inappropriate to rely on the market to create fair and adequate access to public user data on the Internet. Section five contains a conclusion and suggestions for the future.

1. Public User Data on Platform Websites

The World Wide Web consists of hyperlinked websites that display text, images, and other digital media and information on a web browser. As the Internet expanded, platform services grew rapidly to facilitate information exchange between Internet users. Platforms such as Facebook, Twitter, Instagram, LinkedIn, Google, Craigslist, Yelp, and Airbnb provide popular web services that allow people and

businesses to create, upload, search, and/or share user-generated content. Platform services improve the usability of the Internet for ordinary users who lack programming skills. An ordinary Internet user can rely on platform services to share multimedia files, to search and exchange information, and to communicate with people globally without understanding technological details that enable activities on the Internet. A large portion of Internet users today rely on the tools and services offered by platforms to communicate and share information with others (Reyman, 2013, p. 513).

Depending on the nature of a platform's business, some or all of user-generated content hosted by the platform may be available for public access. Platform users can determine which user-generated content is made publicly accessible by examining a platform's policy, such as terms of service or user agreement, and also through their own interactions on and off a platform. Moreover, a platform may offer privacy settings for user-generated content, which allow users to specify how broadly their content may be shared. For example, according to Facebook's data policy, users' content is viewed by anyone if it is published under the "Public" setting, including "people off of Facebook and people who use different media . . . and other sites on the Internet".[1] It also notes that some information shared on Facebook is always made publicly available, such as some information under user profiles. Websites such as Yelp, TripAdvisor, Airbnb, and Kijiji that publish user reviews or user-created ads will make most of the user contributions freely accessible to the public to maximize traffic to their website and to facilitate business.

Public user data on platforms, such as user profiles, schedules, time stamps, preferences, locational data, or historical data, are valuable resources that can be used for various commercial and non-commercial endeavours. Third-party use of such data can contribute to economic growth and development in society. Possible third-party commercial applications include new services that create access to aggregated web data (e.g., price aggregators), analyze data (e.g., personalized ads), or offer new web tools (e.g., mapping locational data to create a visual display) (Scassa, 2017; Hirschey, 2014; Din, 2016; Gladstone, 2001). For example, in *hiQ Labs v. LinkedIn* (2017), hiQ Labs scraped publicly accessible user profiles on LinkedIn and sold the

[1] Consulted August 2021, from https://is-is.facebook.com/help/203805466323736.

statistical analysis of the data to businesses who wanted to learn more about their employees' skills.

Also, public user data could be used for various non-commercial purposes, such as research, lawmaking, and education. For example, public user data may be used to study human behaviour and societal issues (Landers et al., 2016). Governments may require access to public user data for regulation and planning purposes (Scassa, 2017). Moreover, web data can facilitate artificial intelligence and machine learning research (Mavridis, 2011; Pozzi et al., 2016; McClelland, 2017).

In spite of the many possible uses for public user data, there is considerable confusion in law about the accessibility of public user data on platform websites for third-party use and about what restrictions platform companies can impose on third-party data scraping. Third parties who seek to reuse public user data may need to bypass several technological access barriers, and they are subject to numerous laws which can vary by jurisdiction. Privacy laws, such as the *Privacy Act* of Canada, the *Personal Information Protection and Electronic Documents Act* (PIPEDA) of Canada, and the European Union's *General Data Protection Regulation* (GDPR), apply to the processing of personal information or information about an identifiable individual by government institutions and the private sector. Also, third parties cannot use copyright, trademark, and other intellectual property law-protected content without permission from intellectual property owners (subject to exceptions in law; see Section 3.2). Moreover, there are multiple laws, including contract law, tort law, copyright's anti-circumvention provisions, that platform companies can enforce against third parties accessing platform websites to collect data. The core discussion in this chapter will concern the relationship between platform companies and third-party data scrapers.

Privacy law issues are complex, and I will examine them briefly here. Privacy law addresses data regulation aimed at enhancing the digital economy while protecting individuals' right to their personal data. Both platform companies and commercial data scrapers are subject to numerous duties under the privacy laws of relevant jurisdictions to lawfully, fairly, and transparently use personal data, including the core requirement of obtaining meaningful consent of individuals where appropriate when commercial users collect, use, and disclose personal data (GDPR, Article 5(1)(a); PIPEDA, s. 6.1 and schedule 1; GoC, 2018, pp. 2–3). Privacy laws vary by jurisdiction on their treatment of publicly accessible data. For instance, Canada's PIPEDA

includes exceptions to the requirement of obtaining meaningful consent when collecting, using, or disclosing some types of publicly available personal information, such as personal information that appears in a public telephone directory, in public business directories, and in printed or electronic publications (PIPEDA, s. 7; *Regulations Specifying Publicly Available Information*). However, data scrapers may not rely on a broad interpretation of such exceptions for consent to use personal web data because the Government of Canada has recently acknowledged that individual posting of personal information on a public website can attract privacy interests and there should not be unconstrained access to such data (ETHI, 2018, pp. 27–28; GoC, 2018, p. 3). On the other hand, the US district court in *hiQ Labs v. LinkedIn* (2017) noted that LinkedIn users' expectation of privacy on publicly posted user profiles is uncertain at best in light of LinkedIn's inadequate protection of its members' privacy interests, including allowing third-party access to such data without users' knowledge or consent. Canada and other countries are making efforts to update data protection laws to build a strong, coherent data protection regime in light of emerging information technology. These efforts are timely considering the recent privacy scandals relating to a popular social networking website (see Anderson, 2018).

There are various legal and technological measures that can discourage or bar third-party data scraping. These measures provide security against unauthorized access to a website. Platforms can also strategically use these measures for proprietary management of user-generated content and to maximize profits. For example, LinkedIn filed a lawsuit in 2016 in the United States against multiple anonymous data scrapers for automatically scraping LinkedIn user data. LinkedIn claimed that these unknown scrapers violated the US *Computer Fraud and Abuse Act* (CFAA), section 1201 of the *Digital Millennium Copyright Act* (DMCA), state criminal law on unauthorized computer access and fraud, breach of contract, trespass to chattel, and misappropriation (*LinkedIn Corporation v. Does*, 2016). Threats of litigation from platforms with multiple claims can create a chilling effect on third-party use of public user data.

Platform companies are commercially motivated; hence, the goal of profit maximization can overrule fair and transparent regulation of user-generated content. Platforms often generate revenues by including ads on their websites for human users and selling user data to their business partners and advertisers (Hirschey, 2014,

pp. 898–899; DeNardis, 2014, p. 155; *Douez v. Facebook Inc*., 2017). The business model based on advertising profit encourages platforms to tightly control third-party access to and use of the hosted user data because ad profits increase with more users and user-generated content on the website. It is commercially advantageous to establish themselves as the only access portal to the large user-generated content. By tightly controlling the hosted content, platforms can sustain users and limit competition. Hence, popular platform companies may be willing to use their large resources to discourage third-party data scraping that appears detrimental to business (e.g., *Facebook v. Power Ventures*, 2009). Such business tactics privatize user-generated content, including factual data, and cause the public to miss out on possible innovation and new knowledge in society.

Governments should expand their data regulation on the Internet to improve access to publicly accessible user data. In the regulation of public sector data and personal data, some governments have recognized data as "an innovation currency" and "the lifeblood of the knowledge economy" as it is vital to economic and social progress in an information society (EC, 2011, p. 3; GoC, 2018, pp. 1–2). Businesses are primarily driven to maximize profit (Lemley & Lessig, 2001, p. 11); they cannot be relied on to make fair choices about valuable resources such as public user data or to prioritize public interest over private commercial benefit. Third parties that fairly and transparently use public user data should not have to negotiate with platform companies to access such data. After all, search engines routinely access and collect data from publicly accessible areas of platforms and other websites (Christian, 2017). Lawmakers will need to examine laws in multiple areas to improve third-party use of publicly accessible facts on the Internet.

2. Legal Intervention in the Development of the Internet

Although technology is often solely credited for the World Wide Web, the law was also important for creating the Internet as a non-discriminating open network that facilitates information-sharing worldwide and permits anyone to freely contribute data and technology to it (Lemley & Lessig, 2001). The Internet's tremendous capacity to encourage economic and social benefit is closely tied to its underlying architecture reflecting open access ideology (Zittrain, 2008). Platform services that impose excessive and unfair restrictions on the use of

public user data threaten the open nature of the Internet and the Internet as a generative technology. As this technology evolves, lawmakers must continue to play an important role in protecting the freedom of data and information online.

The earliest version of the Internet was built on top of existing telephone networks. Lemley & Lessig (2001, pp. 11–13) note that the Internet would not have evolved into a generative technology without the innovation policy that transformed the telephone networks from a monopolized resource to a general-purpose network (also see Zittrain, 2008, pp. 21–22). In the 1950s and 1960s in the United States, disputes over the use of third-party attachments on American Telephone & Telegraph (AT&T) networks led to decisions that opened up the telephone networks to work with third-party inventions. In *Hush-a-Phone Corporation v. U.S.* (1956), it was held that a user-invented cup-like device that attached onto a telephone receiver to allow a private conversation could not be barred by the telephone company because there was no evidence to support that its use impaired the telephone system or created public injury. The US Federal Communications Commission in *Use of the Carterfone* (1968) also rejected the telephone company's argument for absolute control over the telephone networks, and held in favour of allowing a third-party device of a two-way radio to be attached to the telephone system as long as the device did not adversely affect it (Wu, 2007). These decisions introduced a new innovation policy in law, turning the telephone networks in the United States into an open resource for inventors to build innovations that could address heterogeneous user needs. The inventors' freedom to access the physical layer of the telephone network at any point along the network (rather than access being granted at the discretion of the telephone company) made the physical layer generative. The decisions paved the way for inventions, such as fax machines, answering machines, and modems. Moreover, it became possible for academic researchers and amateurs to design and build the Internet on top of telephone networks. This attribute of the underlying network also influenced the development of the Internet as an open network and a generative technology that encourages users to contribute data and innovation without discrimination (Zittrain, 2008, pp. 22–35). Businesses generate profit by blocking potential competitors' access to the details of their products. However, the Internet was initially built by academics and amateurs who embraced open access and

information sharing rather than being motivated by profit-seeking. Their decentralized collaboration led to the Internet as we know it today (i.e., the World Wide Web), which allows anyone to access and add data and services to the Internet from anywhere in the world, rather than a technology that is centrally located and controlled by a private entity (Saltzer et al., 1984; von Hippel, 2005, Chap. 3). This design allowed information and communications technology to advance rapidly.

Today, popular platform services improve users' access to information and communication over the Internet. Ordinary citizens often depend on these platforms for online communication (*Douez v. Facebook*, 2017; Reyman, 2013, p. 513). However, platform companies should not be allowed to privatize the massive amount of data and information contributed and generated by platform users. Popular platform websites can seriously challenge data and information freedom on the Internet and hinder technological progress. Instead, lawmakers must continue to play an important role in shaping the Internet, including the use of publicly accessible data to maximize public benefit.

Furthermore, third-party access to public user data will be necessary as the Internet evolution enters the next phase, which may be characterized by a proliferation of automated intelligent programs (i.e., bots) that deliver information and services to users (Berners-Lee et al., 2001). This next phase will depend on technology such as semantic web, artificial intelligence, and machine learning, which must process large amounts of data to extract new information or to create useful services (Mavridis, 2011; McClelland, 2017). Third parties need automated access to existing web data for these technologies to evolve.

3. Access Barriers to Public User Data on Platforms

Platform companies can use various technological and legal tools to bar unwanted third-party bots from accessing websites and gathering user-generated content. In this discussion, data-scraping bots are programs that enter target websites to collect publicly accessible data. Search engines widely use such programs to gather information about websites (e.g., Google's Googlebot). Bot users are individuals who use these programs to gather data from someone else's website. Data-scraping bots in this discussion are not malicious programs

designed to purposefully harm websites or change or delete data from websites.

3.1 Technological Barriers

For Internet users, the main tool for accessing web content is a web browser. Nonetheless, browsers typically do not offer a means to reuse web data; they do not allow users to locate and save a large amount of web data into an easy-to-access format for future use (WebHarvy, n.d.). Users have three options for collecting web data: manual data scraping, downloading data from an application programming interface (API) if available on target websites, or automated data scraping.

To manually scrape data, users must locate and copy data on one or more web pages and then clean up, convert and save relevant portions into a particular format and/or a database for future use. This process can be extremely laborious if the user wants to extract frequently updated data or a large dataset from one or more websites. A website usually has multiple web pages. Manual data scraping is inefficient if the target website(s) is constantly updated and expanded.

Alternatively, some websites offer their data in a structured format for third-party use over an API. If so, data users can sign up to an API and download web data in an easy-to-use format. However, this method of data sharing may not be fair or transparent because it allows websites to control what data, when, and how much data are shared with third parties (Hirschey, 2014, p. 906). The data available through APIs may not match the latest data displayed or used on target websites, and some data may not be available at all via APIs when websites want to avoid third-party analysis.

Lastly, automated data scraping uses a bot (i.e., a program) to gather web data directly from target websites. A bot can access the latest data published on a website at the time of scraping, which is what a human user would see on a browser. As noted above, data scraping is a labour-intensive process. Automated data-scraping technology is an efficient tool because a bot can scrape publicly accessible web data significantly faster and more thoroughly than human users. However, automated data scraping may be difficult when bot users have not obtained permission to access a website that uses complex anti-bot technologies to stop bots from entering the website. Without firm, enforceable rules that regulate automated data scraping (and consequently, the use of automated data scraping and anti-bot technologies), public user data may have limited subsequent use.

The process of automated data scraping can be broken down into three steps (Peterson & Davie, 2000, pp. 640–645; ScrapeHero, 2014; Alhenshiri, 2012). First, a hypertext transfer protocol (HTTP) request is sent from the third party's machine to a platform website's web server. An HTTP request is a request to access the web page associated with a URL (i.e., website address). The web server sends the requested web page to the third party's machine in response to this request. This step occurs whether the request is made from a browser of a human user or a data-scraping bot. The fetched web page usually consists of hypertext markup language (HTML), codes, metadata (i.e., additional information about the web page), and contents displayed on the web page, such as images, texts, and web links. The second step involves parsing and cleaning up the fetched web page. This step is necessary because machines cannot interpret the contents of a web page like a human reader. Thus, a data-scraping program examines the fetched page, discards any unimportant parts, and keeps relevant data and web links to other web pages of the website. The third step involves storing the extracted data in a desired format and/or a database for future use. A bot will repeat these steps until there are no more web pages to visit on a website.

Unlike human users, bot users have to overcome the technological challenges of entering a website. A website is a black box to everyone but its owner. It is unclear from the outside how a website monitors and polices website users. Hence, a data-scraping bot usually needs to be programmed specifically to perform on a target website to fetch relevant data against the website's layout, structure, and technological access barriers. For this reason, one strategy to discourage third-party data scraping is to regularly change the website's layout and structure to throw off bots. A data-scraping bot that requests access to a website can encounter multiple anti-bot technological protection measures (TPMs) that discourage or stop automated access to a website, such as a login requirement, captcha tests, cookies, scripts, and IP blocking (Kerr, 2016, pp. 1161–1170). For example, some TPMs on a website operate inconspicuously for human users, such as session cookies, scripts, and networking tools that track and monitor visitors' browsing patterns. Websites can analyze this information to identify bots from human users and block only bot users. A website can refuse a bot's access, for instance, by blocking the IP address associated with the bot user and ignoring any HTTP requests from the blocked IP address (i.e., IP blocking). Thus, unlike human users, a bot may need

to change its IP address multiple times or change its login information to visit publicly accessible areas of a website.

Bot users who want to examine data from a large website or multiple websites will likely need to overcome numerous technological barriers to access a website. Some TPMs are trivial (e.g., captcha and login requirements), and some are complex technology that can be difficult to bypass to enter a website. Increasingly sophisticated anti-bot technologies are significant access barriers to bot users (especially if Internet users have small resources to access bot technology) and deter third-party use of public web data (see Sawatzky, 2015). Popular platform companies with large resources can implement a combination of technological access barriers to discourage and block third parties from examining their web data. Circumventing TPMs, whether trivial or complex, can also raise legal consequences for bot users (*Facebook v. Power Ventures*, 2009; *Craigslist v. 3Taps*, 2013; also see the following section).

Third parties using data-scraping bots have two choices absent enforceable rules on using data-scraping technology: try to avoid detection by websites and data scrape discretely or convince the target websites to permit bot access for automated data gathering. Without enforceable rules to rely on, data scrapers may prefer to avoid detection by websites to avoid conflict. Risking detection of their bot can lead to punishment (e.g., website access denied) and lawsuits from target websites. Popular platforms have the financial resources to threaten lawsuits and engage in lengthy litigation with data scrapers. Avoiding detection may be more than a practical choice for a data scraper because it is difficult to predict how a website will react to third-party data scraping. If data scrapers communicate with platform services before collecting publicly accessible web data, platforms can identify bot users and selectively block their activities (see Scassa, 2018). For example, in exchange for granting bots' access, platforms can require data scrapers to agree not to publish critical or undesirable information about their business or non-participation in related business.

The robots exclusion protocol (REP) is a method that allows websites to specify the rules of automated access and use, and bot users voluntarily follow them (Lundblad, 2007). Websites can implement the REP by including a file called robots.txt in the root directory, which has a set of instructions for bots that request access. The file contains information, such as which bots are allowed to crawl the

website, which portions of the website can be crawled, and how fast bots can fetch data from the website. A bot can be programmed to ignore the robots.txt file on a website, but programmers generally encourage each other to follow it out of good faith (Alhenshiri, 2012). However, websites can also include instructions that unfairly treat some third-party bots and refuse their automated access to content that is publicly accessible on a browser. The REP cannot prevent disruptive third-party access to websites or protect third parties' automated access to publicly accessible user data.

As noted above, the Internet, as an open network and a generative technology, has a tremendous potential to encourage innovation and progress. Nonetheless, a small number of companies (e.g., Google, Microsoft, and Facebook) dominate the big tech and Internet business. A few businesses or an oligopoly should not control valuable innovation resources, such as public user data. Moreover, the costs to sidestep anti-bot technologies will likely increase over time as technology evolves. Without proper regulation of data scraping, it can be quite inefficient and costly for third-party data scrapers to access publicly accessible data on large websites. When third-party data scraping is performed without causing harm to individuals or target websites (i.e., it is carried out politely by fetching public user data from publicly accessible web pages without significantly interfering with the website's operation), firm rules or law should support third-party access to data and deter platform companies' active interference.

3.2 Legal Barriers

Platform companies can also rely on multiple legal measures to deter third-party data scraping. Depending on the jurisdiction, platforms can bring lawsuits against unauthorized data scrapers for violating multiple laws, including contract law, tort law, copyright law and its anti-circumvention provisions, and criminal laws prohibiting access to a computer system (Snell, 2016; *LinkedIn Corporation v. Does*, 2016; Scassa, 2017, 2018). However, policy-makers can support open data and the open access ideology on the web by reviewing and modernizing appropriate areas of law to encourage third-party use of public user data. The discussion in this section will primarily be based on the laws of Canada.

Platform companies can bring a breach of contract claim against data scrapers. Platform users are bound by the website's terms of use

or user agreement, which is enforced in contract law (*Century 21 v. Rogers*, 2011; *Trader v. CarGuru*, 2017). Some platform websites' user agreements may contain provisions that prohibit data scraping. Broad anti-data-scraping provisions protect platform companies' investment and future profits. However, such practice does not recognize the public's interest in third-party use of public user data. Furthermore, broad anti-data-scraping provisions contradict the Internet as a generative technology.

For example, according to Facebook's terms of service, Facebook does not allow automated data collection unless Facebook preauthorizes it. Moreover, LinkedIn's user agreement does not allow users to "[d]evelop, support or use software, devices, scripts, robots, or any other means or processes (including crawlers, browser plugins and add-ons, or any other technology or manual work) to scrape the Services or otherwise copy profiles and other data from the Services." Platforms can discourage undesirable third-party data scraping by threatening lawsuits for violating the terms of use.

In contract law, online user agreements may become binding on a user when the user acknowledges the agreement by clicking on a box labelled "I agree" at login or website registration (i.e., a click-wrap agreement) (*Century 21 v. Rogers*, 2011; *Trader v. CarGuru*, 2017; *Douez v. Facebook*, 2017). In some cases, the act of using a website can bind website users to its user agreement (i.e., a browse-wrap agreement). When a bot enters a platform website to gather data, the person running the bot is likely bound by the website's user agreement because bot users typically need to visit the website before running the program to customize it to work against the target website's layout and structure. Contract law presumes that contracts are struck in a free market economy between parties freely entered into an agreement (McCamus, 2012). However, as noted above, when Internet users rely on popular platform services to access information and communicate with others, users cannot reject these platforms and their user agreements.

Platform owners motivated by advertising profit will protect and sometimes even expand their right to control user-generated content on platform websites. Therefore, platforms unilaterally modify user agreements from time to time to reflect any changes in law or business strategy. For example, Craigslist briefly unilaterally changed its terms of use in 2012 to stipulate that Craigslist had exclusive copyright licensing of user-submitted ads on the website, which would grant the company the right to block anyone from using the ads

(Carrier, 2013, p. 773; *Craigslist v. 3Taps*, 2013). Therefore, data scrapers who regularly collect data from a website also need to routinely examine the terms of use for any changes on data scraping. Third-party data scrapers should not presume that access to a website will be allowed on an ongoing basis.

Furthermore, platforms may bring a claim of tort of trespass to chattels against data scrapers (*Century 21v. Rogers*, 2011, para. 285). To make out this claim, platforms must show that a data scraper trespassed on personal property (i.e., web servers) within their possession. For example, there is no possession if a platform runs its website on a third-party server. Platforms must also show that data scraping interfered with their possession of the personal property; that is, they must have suffered some damage as a result of data scraping. Nonetheless, data scrapers need target websites to be functional and to be able to service users to generate user content (see Alhenshiri, 2012). They cannot scrape web data if their bots disrupt or damage the target websites' servers. Still, some US courts have adopted a flexible view on what is sufficient damage to allow this tort claim to be brought against a data scraper, such as data scraping that devalues a website's investment (Din, 2016, p. 438). While the availability of this claim in Canada is uncertain in the context of data scraping, it seems to be a viable claim against data scrapers in some American states (Scassa, 2018, pp. 47–49).

Moreover, platform companies can bring multiple claims against data scrapers under copyright law. Third parties collecting user-generated content from a platform website can infringe the platform's copyright in the collection or compilation of hosted user data or copyright in its website. Copyright law protects against unauthorized copying of original literary and artistic works fixed in a tangible medium, such as photos and written expressions (see *Copyright Act*, ss. 2 & 5(1)). Copyright law does not protect facts or mere ideas (*CCH v. LSUC*, 2004, para. 15). Therefore, third parties are free to use public user data that are facts. However, copyright law provides separate protection for a compilation of data (Vaver, 2011, p. 92; Scassa, 2017, p. 1050; Scassa, 2018, pp. 28–31). There is no separate database protection law in Canada like the European Union's *Database Directive*. Hence, factual data are unprotected, but an original selection or arrangement of facts is protected in copyright law as a compilation (*Feist Publications v. Rural Telephone*, 1991, para. 44). An original compilation can also consist of facts and other copyrighted works. Any substantial use of a

compilation is a copyright infringement (Vaver, 2011, p. 185). A platform must establish that a compilation is "original" under copyright law, which may not be difficult in Canada (Vaver, 2011, p. 101; *CCH v. LSUC*, 2004, para. 34; Scassa, 2018, p. 28). On the other hand, the US Supreme Court in *Feist Publications v. Rural Telephone* (1991, paras. 17–18) stated that since facts are unprotected in copyright law, the protection of compilations of facts in copyright law is "thin."

Data scrapers can also infringe a platform's copyright in its web page (which is a compilation of data and other copyrighted works) when they make a temporary copy of a web page onto their computer to process and extract relevant data (Vaver, 2011, p. 163). Although this step is unavoidable in digital processing, some litigants in the United States have successfully argued that there is copyright infringement when a temporary cache copy of a web page is created on a third-party computer for the purpose of extracting data on the page (e.g., *Facebook v. Power Ventures*, 2009). On the other hand, several US courts have held that unauthorized copying of large amounts of copyrighted works for text- and data-mining analysis falls under the fair use exception and is not copyright infringement (Cox, 2015). When a temporary copy of a web page is made to digitally extract facts or ideas, copyright law should not interfere with third parties' right to use facts and ideas.

Bots must also create temporary copies of a web page to deliver automated services on the Internet (De Beer & Fewer, 2015, para. 5). The United Kingdom adopted a statutory exception in copyright law in 2014, which exempts copies made from lawfully accessed works for text and data analysis in non-commercial research (see UK *Copyright, Designs and Patents Act 1988*, s. 29A). However, this exception does not encourage a variety of third-party web data use described above because it requires data scrapers to obtain permission from target websites before accessing them, and it only exempts non-commercial research use.

There are provisions in copyright law that exempt some unauthorized copying from infringement, such as US fair use or Canadian fair dealing exceptions. Data scrapers making temporary copies of a web page for private study, research, commentary or review, news reporting, or education may rely on the fair use or fair dealing exception (see *Copyright Act*, ss. 29, 29.1, 29.2; Scassa, 2018, pp. 33–41; Aufderheide, 2011). Courts decide whether such uses are fair on a case-by-case basis by weighing several factors. It can be more difficult

for commercial users than non-commercial users to argue fair use or fair dealing. A copyright user who directly competes in the market with the copyright owner will have a harder time arguing fair use or fair dealing (*CCH v. LSUC*, 2004, para. 59). Data scrapers cannot rely on this exception if they waive their fair use or fair dealing rights in a binding contract with a platform service (Cox, 2015, p. 1).

Anti-circumvention provisions in copyright law pose a serious threat to data scrapers. These provisions can bar data scrapers from gathering public user data regardless of the purpose of use if scrapers circumvent a TPM to access a website. The fair dealing exception does not extend to circumventing a TPM in Canada (Scassa, 2018, p. 42). Anticircumvention law is problematic in data scraping because it grants too much power to platform companies to restrict third-party access to web data, including publicly accessible user data. These provisions prohibit copyright users from circumventing TPMs that are intended to limit access to and use of copyright-protected works. Section 41.1 of the Canadian *Copyright Act* prohibits circumvention of a TPM (i.e., any effective technology, device, or component) that is placed to control access to a work. Copyright users cannot engage in actions such as "to descramble a scrambled work or decrypt an encrypted work or to otherwise avoid, bypass, remove, deactivate or impair the TPM, unless it is done with the authority of the copyright owner." A platform's copyright protection in its website allows the platform to enforce anti-circumvention provisions against data scrapers. A wide variety of anti-bot measures on a platform website may be deemed TPMs in Canada because the term "TPM" is broadly defined in Canadian law (*Nintendo America v. King*, 2017, paras. 81–84; Scassa, 2018, pp. 42–43). TPMs discussed in the previous section are likely protected in Canadian anti-circumvention law, and bypassing these measures without authorization can attract liability for data scrapers. For example, programming data scraping bots to change IP addresses to avoid the platform's IP blocking may be considered bypassing a TPM under anti-circumvention law (*Facebook v. Power Ventures*, 2009; *Craigslist v. 3Taps*, 2013).

Copyright law and its anti-circumvention provisions in Canada do not properly balance the rights of the public to benefit from third-party data use against the rights of platform companies. The law requires third-party data scrapers to explain their actions to powerful platform companies to get permission to access target websites. However, as noted above, these are businesses with no duty to

maximize or prioritize the public's benefit from hosted user data. The law does not even allow data scrapers to defend bypassing TPMs as necessary for fair use or fair dealing. Moreover, since platform companies can unilaterally modify user agreements and technological measures on their website to enhance their control of user-generated content, copyright and anti-circumvention laws should not apply strictly against bot users who access a website to examine publicly accessible user data.

4. Discussion

There must be legal intervention to create better access to public user data shared on platform websites. Platform companies can discourage data scraping by increasing anti-bot measures that block automated access and data collection. The possibility of attracting multiple legal liabilities from data scraping can also discourage economically and socially beneficial uses of public user data. Moreover, in *hiQ Labs v. LinkedIn* (2017), the US Northern District Court of California acknowledged that "conferring on private entities such as LinkedIn, the blanket authority to block viewers from accessing information publicly available on its website for any reason [. . .] could pose an ominous threat to public discourse and the free flow of information promised by the Internet." Therefore, the rights of platform companies to create profit must be balanced against the public's right to benefit from third-party data use.

Clearly, both platform companies and data scrapers should be mindful of how their actions affect the general public and the functioning of the Internet. Both parties should exercise care in order to avoid causing harm to each other. One reason why a platform website may refuse an unfamiliar data-scraping bot from accessing and gathering public user data is because there is a possibility that a bot might interfere with the operation of the website. For instance, unlike human users who visit one web page at a time, bots can rapidly and concurrently send the request to visit a website's multiple web pages. Bots' rapid and concurrent access requests can tie up a website's servers, preventing the website from servicing other users. Thus, data scraping should never be done too rapidly to avoid exhausting a website's server resources and disabling the website (Alhenshiri, 2012). Such third-party access can be mistaken as a denial-of-service attack. Websites can use law and technology to block harmful uses of their

resources. However, as noted above, when data scrapers need to gather public user data for commercial or non-commercial purposes, there is usually no incentive to harm or to interfere with the host platform's operations because data scrapers need to retain ongoing access to the website to collect the data.

Thus, it is generally recommended that programmers should develop a well-behaving and respectful bot that does not impose an excessive burden on a platform's web servers. For example, data scraping bots can request a web page from a web server at a similar rate to human users browsing a website (i.e., two to five seconds between each request for a web page) or mimic search engines that crawl the Internet (Sangaline, 2017). Data scrapers can also explain their bot use to target websites by attaching additional information in the HTTP request (Alhenshiri, 2012). If a bot politely enters publicly accessible portions of a platform website without imposing an excessive burden on its web servers, platforms should grant access.

On the other hand, platform companies may have strong incentives to privatize user-generated data and to block third-party data scraping, such as excluding competition and speech that can negatively impact their business. Hence, society cannot depend on platform companies to decide what kind of third-party data use is appropriate. Businesses cannot be expected to promote society's welfare before their other goals (Lemley & Lessig, 2001, p. 11). Businesses exist to generate profit, and can engage in selfish behaviours. After establishing themselves as industry leaders, popular platforms can use their market position and influence to control user-generated content more aggressively to maximize profit, reduce competition, and control speeches about their business. It also harms data and information freedom in cyberspace and the Internet as a generative technology when platform companies use technological and legal measures to discriminate against some third-party data users. For example, most websites welcome automated access by popular search engines, even when some of them commercially use scraped web data, because search engines benefit a website's business by directing more users to it. Google's automated program (i.e., the Googlebot) crawls most of the Internet to build an index for its search engine and uses the fetched content from various websites to provide services like Google News (Christian, 2017). Also, other large online companies may offer partnerships and other commercial incentives to gain access to a platform's user-generated content. However, platforms may be reluctant

to provide access to third parties who do not offer a business advantage.

Lawmakers cannot expect the market to fix platform companies that behave badly. Online platforms should not have free rein over user-generated content because it can be difficult to replace the handful of popular services that control the digital environment. Popular platform companies may have the first-mover advantage (Burstein, 2012, p. 217) and the benefits of the network effects that accumulate over time (Helberger et al., 2015). These factors, coupled with many users' resistance to change and adapt to a new digital environment, allow popular platforms to maintain their positions of power in cyberspace. Allowing platform companies to determine who can use publicly accessible user data (i.e., by retaining laws that require third parties to seek prior permission from platforms to access data) can strengthen the existing oligopoly on the Internet and discourage new and disruptive innovation from other innovators.

5. Conclusion and Recommendations

Publicly accessible factual data on platform websites are a significant resource in a knowledge economy. Nonetheless, existing laws that regulate the relationship between platform data hosts and third-party data users may be outdated and uncertain in the context of data scraping. Popular platform companies have legal, technological, and perhaps financial advantages over data scrapers. Lawmakers should deter platform businesses from controlling third-party use of publicly shared user data. Undertaking this development in law is necessary to promote fair and transparent uses of public user data and to protect the Internet as an open and generative technology. Since Internet activities can occur across national borders, follow-on research can consider international guidelines for automated data scraping and web data use.

References

Alhenshiri, A. (2012). *Crawling the web: Creating data indices* [PowerPoint slides]. Dalhousie University. https://web.cs.dal.ca/~anwar/ir/lecturenotes/l13.pdf

Anderson, M. (2018, April 6). *Facebook privacy scandal explained*. CTV News. https://www.ctvnews.ca/sci-tech/facebook-privacy-scandal-explained-1.3874533

Aufderheide, P. (2011). Copyright, fair use, and social networks. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 274–299). Routledge.

Berners-Lee, T., Hendler J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43.

Burstein, M. J. (2012). Exchanging information without intellectual property. *Texas Law Review*, *91*, 227–282.

Carrier, M. A. (2013). Only "scraping" the surface: The copyright hole in the FTC's Google settlement. *UBC Law Review*, *46*(3), 759–790.

Christian, J. (2017, November 22). *We still don't know how Google News works*. The Outline. https://theoutline.com/post/2512/we-still-don-t-know-how-google-news-works?zd=1&zi=xdu635x6

Constantine, J. (2012, August 22). *How big is Facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day*. TechCrunch. https://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/

Cox, K. L. (2015). *Issue brief: Text and data mining and fair use in the United States*. Association of Research Libraries. http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf

Din, M. F. (2016). Breaching and entering: When data scraping should be a federal computer hacking crime. *Brooklyn Law Review*, *81*(2), 405–440.

De Beer, J., & Fewer, D. (2015). *35918 CBC v. SODRAC: Factum of the intervener, Samuelson-Glushko Canadian Internet Policy & Public Interest Clinic (CIPPIC)* (SCC Court File No.: 35918). https://cippic.ca/sites/default/files/CIPPIC_Factum-CBC_v_SODRAC.pdf

DeNardis, L. (2014). *The global war for Internet governance*. Yale University Press.

European Commission (EC). (2011). *Open data: An engine for innovation, growth and transparent governance* [Communication]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52011DC0882

Gladstone, J. A. (2001). Data mines and battlefields: Looking at financial aggregators to understand the legal boundaries and ownership rights in the use of personal data. *Journal of Computer & Information Law*, *19*(2), 313–329.

Government of Canada (GoC). (2017). *Open Data 101*. https://open.canada.ca/en/open-data-principles#toc97

Government of Canada (GoC). (2018). *Government response to the twelfth report of the Standing Committee on Access to Information, Privacy and Ethics*. https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-12/response-8512-421-344

Helberger, N., Kleinen-von Königslöw, K., & van der Noll, R. (2015). Regulating the new information intermediaries as gatekeepers of information diversity. *Info*, *17*(6), 50–71.

Hirschey, J. K. (2014). Symbiotic relationships: Pragmatic acceptance of data scraping. *Berkeley Technology Law Journal*, *29*(4), 897–927.

Kerr, O. S. (2016). Norms of computer trespass. *Columbia Law Review*, *116*(4), 1143–1183.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492.

Lemley, M. A., & Lessig, L. (2001). The end of end-to-end: Preserving the architecture of the Internet in the broadband era. *UCLA Law Review*, *48*, 925–972.

Lundblad, N. (2007). e-Exclusion and bot rights: Legal aspects of the robots exclusion standard for public agencies and other public sector bodies with Swedish examples. *First Monday*, *12*(8). http://firstmonday.org/ojs/index.php/fm/article/view/1974/1849

Mavridis, N. (2011). Artificial agents entering social networks. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 291–303). Routledge.

McClelland, C. (2017, December 4). *The difference between artificial intelligence, machine learning, and deep learning.* Medium. https://medium.com/iotforall/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-3aa67bff5991

Peterson, L. L., & Davie, B. S. (2000). *Computer Networks: A System Approach* (2nd ed.). Morgan Kaufmann Publishers.

Pozzi, F. A., Fersini, E., Messina E., & Liu, B. (2016). *Sentiment Analysis in Social Networks*. Elsevier.

Reyman, J. (2013). User data on the social web: Authorship, agency, and appropriation. *College English*, *75*(5), 513–533. http://www.ncte.org/library/NCTEFiles/Resources/Journals/CE/0755-may2013/CE0755User.pdf

Saltzer, J. H., Reed, D. P., & Clark, D. D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems*, *2*(4), 277–288.

Sangaline, E. (2017). *Advanced web scraping: Bypassing "403 Forbidden," captchas, and more.* http://sangaline.com/post/advanced-web-scraping-tutorial/

Sawatzky, K. (2015). *Short-term consequences*. https://shorttermconsequences.wordpress.com/2015/06/20/airbnb-listings-in-vancouver-how-many-what-type-where/

ScrapeHero. (2014). *Webscraping using Python without using large frameworks like Scrapy.* https://www.scrapehero.com/webscraping-using-python-without-using-large-frameworks-like-scrapy/

Scassa, T. (2017). Sharing data in the platform economy: A public interest argument for access to platform data. *UBC Law Review*, *50*(4), 1017–1071.

Scassa, T. (2018). Information law in the platform economy: Ownership, control and reuse of platform data. In D. McKee, F. Makela & T. Scassa (Eds.), *Law and the sharing economy: Regulating online market platforms* (pp. 149–194). University of Ottawa Press.

Snell, J. (2016). *Web scraping in an era of Big Data 2.0.* Bloomberg Law. https://news.bloomberglaw.com/tech-and-telecom-law/web-scraping-in-an-era-of-big-data-20

Standing Committee on Access to Information, Privacy and Ethics (ETHI). (2018). *Towards privacy by design: Review of the Personal Information Protection and Electronic Documents Act* [Report]. House of Commons. http://publications.gc.ca/site/eng/9.852663/publication.html

Vaver, D. (2011). *Intellectual property: Copyright, patents, trademarks*. (2nd ed). Irwin Law.

von Hippel, E. (2005). *Democratizing innovation*. MIT Press.

WebHarvy. (n.d.). *What is web scraping?* https://www.webharvy.com/articles/what-is-web-scraping.html

Wu, T. (2007). Wireless Carterfone. *International Journal of Communication*, *1*, 389–426.

Zittrain, J. (2008). *The future of the Internet*. Yale University Press.

Statutes

Copyright Act, RS C 1985, c C-42.

Personal Information Protection and Electronic Documents Act, S.C. 2000, c. 5 [PIPEDA].

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), [2016] OJ, L119/1 [GDPR].

Regulations Specifying Publicly Available Information, SOR/2001-7.

Case Law

CCH Canadian Ltd. v. Law Society of Upper Canada, 2004 SCC 13, (2004) 1 SCR 339.

Century 21 Canada Ltd. Partnership v. Rogers Communications Inc., 2011 BCSC 1196, (2011) BCJ No 1679.

Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962 (N.D. Cal. 2013).

Douez v. Facebook Inc., 2017 SCC 33, (2017) SCR 751.

Facebook v. Power Ventures, 91 U.S.P.Q. 2d 1430 (2009).

Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991).

hiQ Labs Inc. v. LinkedIn Corp., No. 3:17-cv-03301 (N.D. Cal. 2017).

Hush-a-Phone Corporation v. U.S., 238 F2d 266 (D.C. Cir. 1956).
LinkedIn Corporation v. Does, Case No. 5:16-cv-4463 (N.D. Cal. 2016).
Nintendo America v. King, (2017) FCJ No 253, 2017 FC 246.
Trader Corporation v. CarGurus, Inc., 2017 ONSC 184.
Use of the Carterfone in Message Toll Serv., 13 FCC 2d 420 (1968).

About the Author

Haewon Chung is a computer scientist and a legal scholar. Her doctoral thesis at the University of Ottawa examined patentable knowledge management in the context of amateur-driven, open, and collaborative do-it-yourself biotechnology. She has written about decision-support systems in healthcare, open access to knowledge and science, software patents, and intellectual property law.