.

2016 Student Prize

The Editors of the Computers and Law Journal are pleased to announce that the winner of the 2016 Student Prize is Adrian Agius for the article "Considering legal perspectives and an Australian approach to scraping data from the modern web". Adrian received a prize of $1,000. We are pleased to publish the winning entry below.*

Considering Legal Perspectives and an Australian Approach To Scraping Data From The Modern Web

By Adrian Agius

***Adrian Agius** is* a final year Law and Information Systems student at the University of New South Wales. He works in both technology and law research and with data in law.

BIG - *bad world of* - DATA

Web scraping is a technique that allows for the collection of data from the Internet. Unlike human interpretation of browsers, scraping relies on machine-to-machine interaction to retrieve data from a page. Through considered scripting and software production, 'scrapers' can leverage loops, variables and conditions to intelligently extract information from a webpage. The scalability and relative ease of scraping has also seen it even establish itself as a service for hire.[1]

However, despite the demonstrable power of web scraping, issues pertaining to the legitimacy of the technique have somewhat shrouded its benefit in illegality.[2] The digital landscape has long challenged lawmakers to appropriately balance technological benefit with appropriate safeguards. Scraping in particular, has caused concern in the realms of copyright, attracting the attention of courts worldwide.[3] Others have gone so far as to suggest that the technique should be dealt with as a hacking offence,[4] eliciting criminal sanctions.

This paper will explore the (lack of) treatment of web scraping in Australian jurisdictions, proposing a possible framework for dealing with the issue. In doing so, it will explore overseas attitudes towards data scraping, drawing out approaches that may prove useful within an Australian legal context.

TOO MUCH OF A GOOD THING

In 2014, it was estimated that by 2020, the world will have accumulated 44 zettabytes of data.[5] To draw a comparison, the current figure for total data accumulation sits somewhere between 4-5 zettabytes,[6] meaning that the rise in data production is exponentially increasing. For most, this represents little more than a mark of our technological advancement. However, to the technology community it forewarns of a global issue, a world where data volumes exceed our capacity to continue to effectively store, analyse and secure the information contained therein.

Given the extensive automation of interactions where machine-to-machine is concerned, data gets produced in less insightful ways on a more regular basis. It is estimated by 2020 we will be generating in excess of 400 zettabytes per year.[7] This suggests that in order to gain insight from data sets in the future, entities must work harder to extract data of use.[8] Therefore, it is of little surprise that there is a movement towards the adoption of automated scraping and interpretation techniques.

For the purposes of developing a legislative framework, it is necessary to consider such trends when trying to facilitate web scraping. Doing so will provide for appropriate consideration of future issues stemming from increased data production.

SIMPLICITY OF SCRAPING

At its most basic level, web scraping involves the processing of a webpage to process and extract its data. Methodologies do vary, however web scrapers will extract information in both a specific and generic manner. Contrary to popular belief,[9] web scraping does include web crawling, which is responsible for the generation of indexes contained within search engines.

[*] This article is an extended and revised version of the winning entry.

Web scraping should be viewed as a component of data scraping, which also encompasses techniques that may occur locally (offline), rather than purely over the Internet.[10] It should be noted that when data scraping occurs offline, it generally will attract criminal sanctions, with nations such as Australia and the United States having enacted hacking provisions to deal with unauthorised access to systems.

Despite this, having the capacity to tell the difference between different types of scraping proves quite difficult. Often, the difference may lie in a few lines of code, which requires extremely nuanced consideration by courts in order to both categorise and deal with different instances.

GROUNDED IN COPYRIGHT BUT NOT MUCH ELSE

In Australia, there exist few examples where data scraping has been considered by courts, with specific reference to web scraping non-existent. This can largely be attributed to the reluctance of the legal system to explore the intricacies of scraping, instead choosing to adopt a more generalised approach to the issue. This is particularly problematic because it creates legal uncertainty for those considering the use of scraping tools.

In IceTV Pty Ltd v Nine Network Australia Pty Ltd ('**IceTV Case**'),[11] the High Court of Australia discussed whether protections offered by the Copyright Act 1968 ('**Copyright Act**') applied to databases. The IceTV Case specifically explored what constitutes a substantial part of a data set. The scenario in question involved the appellant, IceTV Pty Ltd, regularly replicating a database maintained by the respondent, Nine Network Australia Pty Ltd. In handing down its judgment, the court noted that although an original work, such as respondent's database, may be attributed to an author, it alone cannot be considered a 'substantial part of the whole work [entire television guide].'[12]

In operation, the method of data comparison and replication undertaken by IceTV Pty Ltd, is comparable to web scraping. The subject matter of the case extended the opportunity for courts to comment on the procedure being employed by the appellant. Instead, the judgment ignored any consideration of such a methodology.

The judgment in the IceTV Case was revisited a year later in *Telstra Corporation Limited v Phone Directories Company Pty Ltd* ('**Telstra Case**').[13] Here the court once again considered whether copyright was vested in a compilation. Telstra Corporation Limited, who were responsible for the production of the White Pages Directory ('**WPD**') and the Yellow Pages Directory ('**YPD**'), argued that the defendant breached copyright in regional copies of each of the directories.[14]

In the Telstra Case, the process of producing each directory was examined by the court. The evidence showed that central to the production and publication of each directory, was the use of a computer program which compiled entries previously collected. The presence of this program in compiling the directory rendered it impossible for Telstra Corporation Limited to demonstrate authorship over the content within the directory.

Upon closer inspection, the Telstra Case also considers factors to do with the nature of a work that can be copyrighted. This is particularly important here because the distinction specifically deals with how copyright can be vested in a database, but not in a work that derives itself from that particular database.[15]

Another case, *Dynamic Supplies Pty Ltd v Tonnex International Pty Ltd* ('**Dynamic Case**'),[16] reaffirmed the principles stated in IceTV. Despite reaching a different conclusion on the facts, the court here dealt with the protection of a compilation that was the direct result of human authorship.

The plaintiff, Dynamic Supplies Pty Ltd prepared a chart that detailed the compatibility of using particular printers with computers. The chart itself was derived from a separated database and arranged in a manner that allowed for easy comparison by customers, that is, in a CSV format.[17] It was alleged that the defendant, Tonnex International Pty Ltd, breached the copyright vested in the chart by reproducing it in a pricing chart, which featured the compatibility index.

The defendant argued that the compatibility index in their pricing chart was uniquely generated. However, evidence revealed that five of the nine columns contained in the pricing chart were copied from the plaintiff's compilation, with 60% of entries in the CSV having been replicated.[18] Although IceTV noted that information like titles and title information were limited in the way that they could be replicated, it did not preclude copyright from being vested in works that were simple in nature.[19] In the Dynamic Case, Justice Yates drew particular attention to this, stating that simplicity need not be a value that negates originality, so long as it does not mask an absence of skill and effort.

UNDERSTANDING AN OVERSEAS PERSPECTIVE

In each of the cases examined above, the court has resigned to the application of copyright law as a mechanism of managing the protection of data. And whilst this is suggestive that there is no current position on web scraping in Australia, it would be extremely naive to only consider copyright law as an appropriate instrument to deal with the act of web scraping. Any appropriate framework requires the careful consideration of the technologies involved. Accordingly, guidance from cases heard in overseas jurisdictions may be of use.

In Field v Google Inc ('**Field Case**'),[20] strong consideration was given to the role of a website operator in preventing web scraping. The case considered the act of web crawling, which Google relies upon to populate its search engine. Google's crawling specifically involved creating a cache of that site so that it may be searched. The plaintiff's website contained published works, whose copyright was allegedly breached when the site was cached.[21]

The *Field Case* is important because it examines the nature of web crawling and the tools that may be used to prevent it. These tools are identical to those prescribed for the prevention of web scraping. Google's method of indexing sites occurs automatically, making use of thousands of 'spiders' that crawl the web looking for new sites, updates to existing sites or site removal.[22] To prevent a bot from indexing a particular site, a webmaster may use a piece of code, usually HTML, to direct a web crawler away from the site.[23]

Perhaps the most common method of preventing web crawling, or scraping for that matter, is to make use of the robots exclusion standard. This involves including a file entitled robot.txt in the root hierarchy of a website. Other methods of prevention include dealing with irregular IP requests, using CAPTCHA forms or embedding code only visible to crawlers or scrapers. Regardless of which method is employed by a website administrator, the intent remains the same, to prevent automated traversal of the site preventing outcomes including indexation and data extraction.

In this case, the plaintiff was aware of Google's mechanism for indexing sites and the ability of robot.txt to prevent this. However, the choice was made not to employ robot.txt, which the court viewed as the plaintiff granting the defendant an implied license to both cache and index the site.[24]

The *Field Case* arguably sets a precedent as to the role of mechanisms such as robot.txt in imputing knowledge. In doing so, it also sets a standard as to the basic Internet literacy required for those willing to operate or host content online. Curiously, the court used Google's web crawling practice and guidelines as the basis by which the plaintiff should operate on. This raises serious doubts as to the impartiality of the court's decision, suggesting private companies have heavy influence over the interpretation of technical definitions.

Another case, Facebook Inc v Power Ventures Inc ('**Facebook Case**'),[25] demonstrated extent to which a webmaster can exercise control over the copyright vested in their site. Here, the defendant was utilising data from Facebook profiles to assist users to generate an aggregate social media account. This was despite Facebook's publishing APIs on how to enable this data capture.

Facebook argued that Power Ventures technology was a breach of the copyright vested in Facebook's webpages.[26] The defendant claimed that Facebook was not the custodian of their user's data, thereby meaning no copyright existed. However, the court accepted Facebook's argument that the means being adopted by Power Ventures involved the caching all the HTML of any given page, prior to extracting user data. The HTML included the structure of Facebook's site which naturally belongs to the company, thereby resulting in a breach of copyright.

Control of data is arguably the most important point to come out of the *Facebook Case*. It is likely that had Power Ventures made use of the APIs provided by Facebook, no issue regarding copyright would have been raised. This in itself causes confusion, suggesting that sites which do not provide APIs have no say as to whether scraping is permitted.

However, perhaps the most telling case to come from an overseas jurisdiction relating to scraping was the 2015 EU decision Ryanair Ltd v PR Aviation BV ('**Ryanair Case**').[27] Heard in the Court of Justice of the European Union ('**CJEU**'), the *Ryanair Case* specifically considered the act of web scraping and the ability of webmasters to protect their sites. The defendant, PR Aviation BV ran a price comparison website, collating content from various sites.

The plaintiff, Ryanair Ltd, is an airline company whose content was scraped by the defendant. In constructing their site, Ryanair employed the use of a terms of service which needed to be agreed to access flight pricing. Unlike other terms of service, Ryanair included critical aspects of their terms in a popup window, which required the user to specifically acknowledge their existence prior to continuing on the site. One of the terms included, forbid the use of 'automated systems or software' to extract data from the website. Ryanair sued the defendant for breach of contract after it was determined they were indeed web scraping airline prices in contravention of their terms.[28]

In response to this suit, PR Aviation argued that the contractual terms imposed by Ryanair could not take effect in light of the European Database Directive ('**Directive**'). Under the *Directive*, eligible databases could be lawfully reused in an insubstantial manner;[29] with any attempt to contractually circumvent this provision being rendered invalid.[30]

The court ruled that the database maintained by Ryanair did not fit the definition of database contained in the *Directive*.[31] Departing from previously understood notions, the CJEU then went on to give effect to the contractual terms imposed by the plaintiff in their terms of use. Accordingly, it was ruled that PR Aviation were bound by the Ryanair's terms, including the requirement not to scrape the site.

The *Ryanair Case* will likely resonate with lawmakers worldwide, mainly because it arose out of an environment not dissimilar to a majority of jurisdictions. That is, an environment devoid of any real legal consideration of web scraping. The decision demonstrates the importance of webmasters including terms of use on their site, as well as the future potential such terms have for the regulation of online web scraping activity.

PROPOSING A LEGAL FRAMEWORK FOR DATA SCRAPING

Although Australia has not specifically considered web scraping in either a judicial or legislative context, existing foundations in copyright, coupled with overseas treatment of the issues provide guidance about what a potential web scraping legal framework should look like.

In truth, it's somewhat advantageous that Australia is yet to establish any concrete methodology in the web scraping space, mainly because the issues considered here are not purely based in law. Instead, an approach championed by both lawmakers and technologists working in conjunction with each other should be favoured.[32]

An effective starting point would be to understand the shortcomings associated with existing Australian cases. Australian law has long neglected the impact of contractual and criminal liability in matters where the use of data is called into question. Cases like IceTV and Telstra offer extensive analysis of the types of data that may be covered by copyright law but fail to offer any other insight.

The issue with relying on copyright law for matters where web scraping is concerned, is that any breach will ultimately turn on what is being scraped, rather than the actual scraping itself. Further, in all the cases considered, no action was taken until the scraped data resurfaced. This very much renders current mechanisms reactionary in nature, which is problematic given the fluidity of data flows on the Internet.

In this regard, giving legal effect to contractual provisions included on websites, offers a more proactive means of dealing with web scraping. As was seen in the *Ryanair Case*, terms of use can be employed as a first line of defence to disincentivise potential scrapers; averting potential breaches of copyright. It may be argued that this offers an inordinate amount of control over data that webmasters may not necessarily own (such as user data). However, lawmakers have through copyright mechanisms already determined what types of data may be protected through use of such terms, dispelling this issue.

In Australia, mandatory requirements already exist for the inclusion of privacy policies on certain websites.[33] Thus, it is likely that implementing scraping terms and conditions into current web environments could occur with relative ease.

Arguably, the proposed framework here favours webmasters, giving them ample opportunity to lock down data in an excessive manner. Strong consideration should be given to ensuring this does not come to fruition. The *Facebook Case* discussed at length the role of APIs in data sharing and the control they can offer hosts in sharing.[34] Whilst mandating the use of APIs for non-protected data may be a costly endeavour, it potentially should be viewed as an option if reliance on self-generated terms become the norm and a protectionist trend emerges.

TOWARDS A MORE INTELLIGENT INTERNET

For the most part, web scraping techniques operate in a manner consistent with improvement Internet services and website functionality. In actual fact, roughly a quarter of all Internet activity derives itself from some form of data scraping.[35] There are however instances, where such scraping occurs in a manner that deprives another party of benefits associated with particular data. It is in these circumstances that we require legal frameworks to step in and thoughtfully consider the impacts of this deprivation.

Australia's current legal climate provides a basis by which we can build an effective approach to legal consideration of web scraping. Through thoughtful consideration, both legal and technical, we are able to create an online environment that reflects what would otherwise be acceptable in society. Failing to do so will inevitably result in an Internet that is devoid of regulation and appropriate data safeguards.

[1] Services such as Webscraper.io offer free and paid plans for extensive data scraping.

[2] Myra F. Din, 'Breaching and Entering: When Data Scraping Should Be a Federal Computer Hacking Crime', 81 Brook. L. Rev. (2015), 406.

[3] Ibid.

[4] Facebook, Inc. v. Power Ventures, Inc. 91 U.S.P.Q.2d 1430.

[5] EMC Digital Universe, 'The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things', 2014 <http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>.

[6] Ibid.

[7] CISCO, 'Cisco Global Cloud Index: Forecast and Methodology, 2014–2019 White Paper', 2014 <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html?CAMPAIGN=GCI+2014&COUNTRY_SITE=us&POSITION=PR&REFERRING_SITE=PR&CREATIVE=PR+to+GCI+WP>.

[8] Ibid.

[9] Boettcher I, 'Automatic Data Collection on the Internet (web scraping), 2015 <http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p6_pap.pdf>.

[10] Ibid.

[11] IceTV Pty Limited v Nine Network Australia Pty Limited [2009] HCA 14 (22 April 2009).

[12] Ibid [6].

[13] Telstra Corporation Limited v Phone Directories Company Pty Ltd [2010] FCAFC 149.

[14] Ibid.

[15] Allens, Focus: Full Court Rules White and Yellow Pages Not 'Original' Literary Works 2010 <http://www.allens.com.au/pubs/ip/foipdec10.htm>.

[16] Tonnex International Pty Ltd v Dynamic Supplies Pty Ltd [2012] FCAFC 162.

[17] Comma separated values format.

[18] Tonnex International Pty Ltd v Dynamic Supplies Pty Ltd [2012] FCAFC 162, [41].

[19] Ibid [80] (Yates J).

[20] Field v. Google, Inc., 412 F.Supp. 2d 1106.

[21] Ibid.

[22] Google, Crawling and Indexing, 2016 <https://www.google.com.au/insidesearch/howsearchworks/crawling-indexing.html>.

[23] Field v. Google, Inc., 412 F.Supp. 2d 1106.

[24] Ibid.

[25] Facebook, Inc. v. Power Ventures, Inc. 91 U.S.P.Q.2d 1430.

[26] Facebook, Inc. v. Power Ventures, Inc. 91 U.S.P.Q.2d 1430.

[27] Ryanair Ltd v PR Aviation BV 2015, Case C-30/14.

[28] Ibid.

[29] European Database Directive, Article 6 & 8.

[30] Ibid, Article 15.

[31] Ibid.

[32] Turton W, The Lawmakers Who Control Your Digital Future Are Clueless About Technology, *Gizmodo,* 2015 <http://gizmodo.com/the-lawmakers-who-control-your-digital-future-are-cluel-1773599908>.

[33] *Privacy Act 1998,* Schedule 1.

[34] Facebook, Inc. v. Power Ventures, Inc. 91 U.S.P.Q.2d 1430.

[35] Myra, above n 2.