Jnl. Intellectual Property Law and Practice Advance Access published December 18, 2008

# Scrapping over data: are the data scrapers' days numbered?

Frank Jennings and John Yates*

When Tim Berners-Lee first proposed the use of hypertext language '*to allow a pool of information to develop which could grow and evolve . . .*',[1] he could not have foreseen that hypertext language would simultaneously be the greatest asset to the exploration of the internet and one of the internet's greatest weaknesses. Over the past decade, software developers have produced increasingly sophisticated programs which exploit hypertext links[2] and provide the 'content poor' with an automated means of mining data from the sites of the 'content rich'—a practice known as 'data scraping'.

Data scraping has no universal definition and covers a number of different methods of obtaining data from a website or database, typically by way of a computer program. It includes 'screen scraping', where the scraper program merely extracts the key data which will appear on the end-users' screen display: the screen scraping program will ignore sections of coding and merely seek to extract plain text from a webpage. It also includes 'web-scraping' (or 'web-harvesting'), which involves the use of a scraper program to extract all the data relating to the underlying structure of the html[3] script used to create that website (and not just the data displayed on the screen). Such programs are often referred to as 'bots', 'webbots', 'crawlers', 'harvesters', or 'spiders'.

The losses may include system overload as a result of massive bot activity, loss of advertisement revenue, loss of control of content, and its subsequent devaluation. However, as yet there are no specific statistics attributed to the losses incurred by website owners through third party data scraping.

With the recent growth of data scraping, it is not surprising that website owners are taking increasingly proactive steps to protect their data and the investment they have made in assembling that data. For example, in June 2008, easyJet confirmed that it had sent out letters of claim to a number of online travel firms demanding that they refrain from scraping pricing information from the easyjet.com website, perceiving that the practice added unfair advertising value to the travel firms' own websites.[4] Not to be outdone, the Irish low-cost carrier Ryanair went further in July 2008 and succeeded in obtaining an injunction to prevent the German company Vtours GmBH presenting Ryanair's flights and timetables for sale to Vtours customers.[5] The airline then launched a similar action in Ireland against Bravofly—a Dutch price comparison site[6]—and has started cancelling flights booked via websites that use data scraping processes.

## Key issues

- Data scraping poses huge threats to the content-rich—loss of valuable information and advertising revenue, and damage to IT systems.
- There is no single 'magic bullet' to prevent data scraping; rather website owners need to be proactive in their fight whether that be by enforcing their IP rights, their contractual terms, or taking technical measures to prevent scraping.
- The scraped and the scrapers should pay particular heed to the risk of breaches of data protection legislation in allowing or performing scraping.

This article reviews the legal issues surrounding data scraping and examines the options open to those whose content has been scraped. Given the paucity of case law and specific legislation in this area, it has been necessary to examine the approaches taken to this problem in other jurisdictions as well as the UK.

Website owners may be able to seek redress against data scrapers by bringing civil claims for breach of confidence, copyright infringement, database right infringement, breach of contract, and trespass. Additionally, it is

[*] Partner and Solicitor, respectively, DMH Stallard LLP. Email: frank.jennings@dmhstallard.com

[1] T Berners-Lee, *Information Management: A proposal*, CERN 1990 (http://www.w3.org/History/1989/proposal.html).

[2] Eg, http://www.iwebminer.com/ (accessed 23 August 2008) for an explanation as to the operation of its data scraping programs.

[3] Hypertext mark-up language.

[4] http://www.theregister.co.uk/2008/06/27/easyjet_travel_sites_warned/ (accessed 23 August 2008).

[5] 'Ryanair wins German screen scraping injunction' http://www.ryanair.com/site/EN/news.php?yr=08&month=jul&story=reg-en-100708 (accessed 23 August 2008).

[6] ibid.

theoretically possible that criminal law may intervene and datascrapers may face prosecution under the Computer Misuse Act 1990 and the Data Protection Act 1998. These are considered further below.

## Breach of confidence

In principle, there is no property right in information itself.[7] It is, however, well-established that the confidential information is capable of protection by virtue of the law of confidentiality. This might arise through a duty under contract or as an equitable obligation. In establishing whether an equitable obligation arises, it is important to take into account the nature of the information, the circumstances in which it was obtained, and notice of its confidentiality.[8]

The protection offered by the doctrine of confidence is not relevant to the majority of website owners since the purpose of most websites is to allow the flow of information. While most website owners might display their privacy policy, this will refer to how the website owners will treat an individual's privacy in relation to the information that the individual submits through use of the website. Nevertheless, there are occasions where confidential information is stored online. Some website owners might store confidential information online behind invisible or hidden links on a website which are not easily found by humans, being placed on a page not linked from the homepage or not easily distinguishable by the human eye by being in the same colour as the background. A bot, not limited in the same ways as humans, might find this information.

Where a website owner is aware that it will be in possession of confidential information—either its own or that of one of its users—it is likely to address this through its terms of use. In such cases, the obligation of confidentiality will arise through contract. For example, a website owner might store confidential information online and introduce simple password protection to restrict access. If the website owner has not ensured that the link to the content is secure and it is subsequently emailed or placed on a chat forum, the information can be accessed without the new user having to enter a password. With the rapid growth of 'software as a service' (SAAS) where a user accesses software through the internet (either on a free or paid-for basis) rather than installing it on his local computer, the user will likely store increasing amounts of his data online and, if the service is successful, the SAAS provider will store massive amounts of data on a single server. Typically, this data will be held securely by the SAAS provider but, where the security used is inadequate or there is a lapse in that security, the confidential information might become accessible by all.

In such a situation, it might be possible to bring an action for breach of confidence with a view to prevent further dissemination of the information.

## Copyright infringement

Website content may comprise a number of separate works that may be subject to copyright protection and which might be attractive to a data scraper. For example, the popular music website NME[9] includes

- layout—a typographical arrangement of published edition;[10]
- news, reviews, and feature articles—literary works;[11]
- photographs—artistic works;[12]
- listings information—which may be protected as a copyright in a database[13] or by database rights;[14]
- flash-coded animations—artistic works in film;[15]
- streamed audio content—sound recordings,[16] musical works,[17] and literary works;
- music video content—sound recordings, musical works, literary work, and dramatic work;
- a database containing this content, again protectable as copyright in a database or by database rights;
- the coding of the webpages themselves, protected as a literary work.

A website owner has the right to prevent a number of activities in relation to his content including copying it, posting it on another website, and modifying it.[18]

Among those who set up the very first websites, there appeared to exist the notion that content placed on websites was not subject to copyright protection. This appears to stem from the early declaration[19] by the founders of the web that they relinquished all IP rights in the architecture of the web and the strong

[7] *Oxford v Moss* [1978] 68 Cr App Rep 183.

[8] RG Toulson and SM Phipps, *Confidentiality* (Sweet & Maxwell, 2006), Chapter 3.

[9] http://www.nme.com/

[10] Copyright, Designs and Patents Act 1988 (CDPA) s 8(1).

[11] CDPA s 3.

[12] CDPA s 4(1)(a).

[13] CDPA s 3(1)(d).

[14] CDPA s 3A(1).

[15] CDPA s 5B(1).

[16] CDPA s 5A(1).

[17] CDPA s 3(1).

[18] CDPA s 16(1).

[19] See http://tenyears-www.web.cern.ch/tenyears-www/

concept of 'public domain' in the USA for works made freely available to all.[20] It did not take long for rights-owners (and their lawyers) to rebut this flawed assumption and to show that the law applies to the internet as to everything else.

It is now accepted that, by placing content on a website, a website owner does not grant a user a licence to perform any of the activities reserved to him as copyright owner, save for a limited licence to make a temporary cached copy to display the content on the user's screen. Sometimes, particularly where a 'print' icon is displayed, the website owner grants a licence to print content. Details of these permissions are usually contained in the terms of use.[21]

> *By placing content on a website, a website owner does not grant a user a licence to perform any of the activities reserved to him as copyright owner, save for a limited licence to make a temporary cached copy to display the content on the user's screen*

By their nature, data scraping programs generally copy data held on the website owner's hosting servers without licence. The scraper often re-presents the scraped information on its own website, usually without referring to the original source. In so doing, the scraper's principle infringing acts are copying of the work[22] and communicating the work to the public.[23] Also, where it applies, the scraper might be infringing the content author's 'paternity right'[24] (by not giving the author a credit) or the author's 'integrity right'[25] (if the data are distorted).

Section 17 CDPA provides that 'copying' of protected literary, dramatic, musical, and artistic works 'includes storing the work in any medium by electronic means'[26] and, in respect of any form of protected work, 'the making of copies that are transient or are incidental to some other use of the work'.[27] Thus, while a website user has the benefit of the exception allowing him to download pages for the purpose of viewing the page,[28] this exception to the copying restriction will not extend beyond merely viewing the page to permit reproduction of the content on the user's own page.

As to the second infringing act of communication to the public, the CDPA states that 'communication to the public' in relation to literary, dramatic, musical or artistic works, sound recordings, films, and broadcasts includes 'the making available to the public of the work by electronic transmission in such a way that members of the public may access it from a place at a time individually chosen by them'.[29] This is the right typically infringed by posting content on a website.

While there may be a lack of English case law relating to copyright infringement and data scraping, a decade ago a Scottish court considered copyright infringement in relation to 'deep-linking', which may be viewed as a primitive form of data scraping. A third party may operate deep-links by coding its page in such a manner that a website user is presented with information which appears to be the linker's own, whereas the linker has merely embedded a hyperlink in its coding to take the user to content on the original data owner's site. The case, *Shetland Times Ltd v Wills*,[30] centred on Wills' use of hyperlinked headlines taken from the Shetland Times website on his own site called the Shetland News. When a user of the defendant's website clicked on the hyperlink, the user bypassed the Shetland Times' homepage and was led directly to the relevant story. In bypassing the claimant's homepage, the claimant argued that it was potentially deprived of advertising revenue. Shetland Times founded its claim on the basis that the headlines available on its website were cable programmes[31] and that the inclusion of headlines by Wills infringed copyright under sections 17 and 20 CDPA.[32] The court granted an interim interdict (the Scottish equivalent of an interim injunction) on the assumption that its claim was well-founded and that the balance of convenience lay in its favour, notwithstanding Wills' assertion the headlines complained of were not original as Shetland Times had expended no skill or labour in creating them.[33]

While this case reached the correct decision from a moral perspective, it provided little certainty that, on a

[20] For example, see http://www.copyright.cornell.edu/public_domain/

[21] For which see Breach of Contract, infra.

[22] CDPA s 17.

[23] CDPA s 20.

[24] CDPA s 77.

[25] CDPA s 80.

[26] CDPA s 17(2).

[27] CDPA s 17(6).

[28] CDPA s 28A.

[29] CDPA s 20 (2)(b)

[30] 1997 SC 316; 1997 SLT 669; 1997 SCLR 160; [1997] EMLR 277; Times, 21 January 1997.

[31] Cable programmes at the time of the claim were capable of protection under s 7 CDPA, repealed by Copyright and Related Rights Regulations 2003/2498 Sch. 2 para 1.

[32] CDPA s 20, as then in force, related to infringement by broadcasting or inclusion in a cable programme rather than communicating to the public.

[33] Relying upon *Ascot Jockey Club Ltd v Simons* [1968] 64 WWR 411.

full trial, a judge would have reached the same conclusion. There was no analysis of whether the defendant breached the Shetland Times' terms of use. Further, the outdated statutory language of that time meant that lawyers had to argue that websites fell within the definition of 'cable programmes', a clumsy approach and one which ignored the (admittedly limited) satellite or radio internet connections at the time.[34]

The outcome of *Shetland Times* is consistent with the ongoing *Copiepresse v Google*[35] dispute before the Belgium courts, if for different reasons. In 2006 Copiepresse, a group representing the interests of several Belgian newspapers, brought a claim against Google alleging that Google News,[36] through its use of bots, had infringed its members' copyright by displaying links to, and caching, their news stories. This deprived them of traffic and, as in *Shetland Times*, of advertising revenue. Copiepresse claimed damages of €49 million. Google countered that, under the E-Commerce Directive,[37] as a provider of 'information society services' it was permitted to make temporary cached copies of webpages and hyperlinks.[38] Further, it argued that it was permitted to incorporate such works by virtue of the Belgian equivalent[39] of the fair use defence for news reporting as provided for in the CDPA.[40] The Court of First Instance rejected these arguments and found in Copiepresse's favour. Google has appealed the decision, which is still outstanding. Meanwhile, Copiepresse has now launched a new claim for damages to the Court of First Instance.[41]

Copiepresse's members could easily have prevented Google's bots from scraping their sites, thus avoiding litigation, by adopting a long-standing protocol to exclude scraping programs.[42] Google suggested that the reason that they did not do so was they wished to retain a presence on Google search results and, unsurprisingly, Google removed links to content of Copiepresse's members. However, these links have since been restored and the Copiepresse members are now using the 'NOARCHIVE' tag to prevent cached material being displayed to end-users.

Google, through its constant innovations and attempt to stay ahead of the competition, often finds itself a subject of court actions. Since its launch in 1998, it has moved from merely displaying hyperlinks to content on the web, to caching, and displaying that content on its own pages and is a leader in search-related advertising.[43] With Google's share of the UK and Australian search market estimated at 90%,[44] any de-listing or lower ranking of a business by Google could damage that business quickly. Therefore, arguably a business will not take action against Google unless it has a serious complaint that its business is being jeopardized. While some internet purists will be horrified to find Google being made a scapegoat (again), often due to the inability or unwillingness of others to adapt or for the law to keep up, for content owners this decision is a welcome step against data scraping. However, this battle between Copiepresse and Google is likely to continue.

## Database rights

The Copyright and Database Right Regulations 1997[45] implemented the Database Directive[46] and effectively introduced dual protection in the UK for databases. The Regulations introduced a definition of 'database'[47] and a new IP right, a *sui generis* right known in the UK as the 'database right' to protect economic investment in databases. The Regulations also clarified that databases are protectable by copyright, provided they meet a new test of originality unfamiliar to English lawyers but common in continental Europe, that of the 'authors' own intellectual creation'.[48] The UK, with its lower 'sweat of the brow' test of originality, previously faced no difficulty in protecting databases as literary works (tables) but this test of the 'authors' own intellectual creation' in continental Europe effectively ruled out widespread copyright protection for databases.

The Regulations state that a person infringes the database right if, without the consent of the owner of the right, he extracts or re-utilizes all, or a substantial part,

[34] Today's technology would be WiFi or WiMax.

[35] Court of First Instance. Brussels No. 06/10.928/C. English translation of judgment available at http://www.copiepresse.be/13-02-07-jugement-en.pdf (accessed 23 August 2008).

[36] Google's news websites' search engine http://news.google.com/

[37] Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

[38] Directive 2000/31/EC Art 13.

[39] Art 22 s 1 of the act of 30 June 1994 relating to copyright law and related rights (1994).

[40] CDPA s 30(2).

[41] Copy of the summons is available at http://copiepresse.be/pdf/summons.pdf (accessed 23 August 2008).

[42] The protocol is known as the 'robot exclusion standard'. For example, see http://www.robotstxt.org/orig.html (accessed 23 August 2008). Additionally, Google will not cache any page that contains the word 'NOARCHIVE' in its code. See http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35306 (accessed 23 August 2008).

[43] http://blogs.wsj.com/biztech/2008/08/14/google-extends-its-online-ad-lead/ (accessed 23 August 2008).

[44] http://www.bizreport.com/2008/07/googles_market_share_nudges_90_down_under.html (accessed 23 August 2008).

[45] SI 1997/3032.

[46] Directive 96/9/EC on the legal protection of databases.

[47] CDPA s 3A(1).

[48] CDPA s 3A (2).

of the contents of the database.[49] 'Extraction', in relation to any contents of a database, means the permanent or temporary transfer of those contents to another medium by any means or in any form.[50] 'Re-utilisation', in relation to any content of a database, means making that content available to the public by any means.[51]

A user who extracts the data from the database and re-utilizes it by putting it on his own website clearly infringes the database right. He will also infringe the copyright in the database[52] if he copies a substantial part and communicates it to the public through his website.

The infringement of database rights by linking was considered in the German case *StepStone v Ofir.*[53] StepStone operated a recruitment website which contained a collection of job advertisements. As in *Shetland Times* Ofir deep-linked to StepStone's website. This allowed Ofir's users to directly access the job offers contained on the StepStone site without visiting the homepages of the StepStone site. The German regional court took a wide view and determined that the advertisements constituted a database and that StepStone was afforded protection in relation to the copying, distribution, and representation of that database by virtue of its database right. The court held that Ofir's deep-linking to parts of the StepStone site constituted unlawful distribution. Thus, Ofir infringed the database right held by StepStone.

*The ability of website owners to assert infringement of database rights by a data scraper or linker depends on the interpretation of 'database right'*

The decision that the linking in *StepStone* is an infringement of database right may be applied in other jurisdictions within the EU. Though, in the French case of *Cadremploi.fr v Keljob*,[54] deep-linking was prohibited, not for infringement of database right, but rather on the ground of unfair competition[55] by the unlawful user.

The ability of website owners to assert infringement of database rights by a data scraper or linker depends on the interpretation of 'database right'. While the German courts took a broad view as to what constituted a database right in *StepStone*, British courts have recently moved towards a narrower construction of the term following a European Court of Justice ('ECJ') ruling that narrowed Mr Justice Laddie's earlier, broad interpretation of it.[56] The ECJ stated:

> The expression 'investment in . . . the obtaining . . . of the contents' of a database in Article 7(1) of the [Database Directive] must be understood to refer to the resources used to seek out existing independent materials and collect them in the database. It does not cover the resources used for the creation of materials which make up the contents of a database.[57]

Thus, while a substantial investment in obtaining, verifying, and presenting the content of a database may be protected by the Database Directive, any investment in creating the content remains unprotected.[58]

Having identified the nature of database rights, it is crucial to identify who may bring a claim. In *Copiepresse*, Copiepresse sought to argue that Google's links in its news search results had infringed its database rights. The court rejected this argument on the ground that Copiepresse, as a management company for copyright works, did not have the requisite standing to bring such a claim. Under Belgian law, it is only the holders of the *sui generis* right or the producer of the database that may bring a claim.[59]

Website owners wishing to assert their database rights against data scrapers are advised to ensure that any investment in the creation or compilation of any database contained with its website is verifiable. The narrowing of protection by the ECJ led to the EU Commission to evaluate the scope of the *sui generis* database right in 2005–2006, but this process has not yet resulted in any changes.[60] Given the ECJ's ruling in *William Hill*, it is unlikely that website owners will be able to include any investment made in creating the original content of the website database as part of an action against a data scraper who unlawfully extracts or uses information from it. The ECJ has made it clear that the creation of the database itself is the only 'investment' that will be deemed to create a database

[49] Regulation 16, SI 1997/3032.

[50] Regulation 12, SI 1997/3032.

[51] Regulation 12, SI 1997/3032.

[52] Provided the database complies with the new higher test of originality mentioned above: the author's own intellectual creation.

[53] LG Köln ZUM 2001, 714.

[54] Tribunal de Grande Instance de Paris, 8 January 2001.

[55] Unfair competition may be equated to the common law cause of action 'passing off' but extends beyond the traditional test of goodwill, misrepresentation, and damage.

[56] *British Horseracing Board Ltd and others v William Hill Organisation Ltd* [2005] All ER (D) 149 (Jul).

[57] Case C-203/02 *British Horseracing Board Ltd and others v William Hill Organisation Ltd.*

[58] C Smith, *Saddling up for the database rights race* (2005) 155 NLJ 28.

[59] B Michaux, *Droit des bases de donnees* (2005 Kluwer) p 166.

[60] http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf (accessed 23 August 2008).

right. With such limited protection affordable under an action founded on a database right, creators of website databases may seek better protection elsewhere.

## Breach of contract

Currently, the majority of websites include terms of use which determine what users are permitted to do in relation to an owner's content. A legal question arises as to whether such terms of use are equally enforceable against users and data scrapers.

While websites and their terms of use may be a distinctly modern phenomenon, case law in this area can be traced back over more than a century. The two key factors at play in the formation of contracts online (and hence website terms of use) are reasonable notice[61] and affirmative assent to terms.[62]

Many website users are familiar with the concept of a 'click-through' or 'click-wrap' licence even if they are not familiar with the names. A user is required to click a tick box next to 'I Agree' or 'I Accept' to confirm assent to certain terms—these for, example, may relate to the website owner's conditions of sale. 'Browse-wrap' licences are a step removed from click wrap licences in that the website owner is not required to take any positive step to acknowledge terms and conditions since browsers are referred to terms and conditions via hyperlink.

These licences have their origins in 'shrink-wrap' licences which began to appear in the late 1980s and accompanied physical copies of computer software. Under a shrink-wrap licence, the prospective licensee can only read and accept the licence after removing the shrink-wrapping from the product purchased.

The enforceability of shrink-wrap licences was called in to question in the USA, where two distinct lines of thought developed. In *Pro-CD v Zeidenberg*,[63] the court held that a shrink-wrap licence was enforceable on the ground that the defendant could read the shrink-wrap licence at his leisure before installing the software but did not reject the software. However, in *Klocek v Gateway*,[64] Mr Klocek successfully argued that Gateway's standard conditions, supplied with a computer, were not enforceable against him and other Gateway users.

It is also US case law which leads the way in relation to click-wrap licences. In *Specht v Netscape Communications Corp*,[65] the court considered requisite consent to contractual term in the context of free software downloads. The court, following *Klocek*, found that Netscape's terms were unenforceable because there was no meeting of the minds before the purported formation of the contract. Absence of consensus was demonstrated by the fact that a user did not have to click on an icon or link to indicate assent before downloading and using the software. The corollary is that, if there is such a facility for assent, the licence may be enforceable. This did not escape the attention of the court, which stated that the assent to a contract may be validly made by a '*click of a computer mouse transmitted across the invisible ether of the internet*'. It did not matter whether the user read the terms before making that click.

*Browse-wrap licences add a further layer of complexity because the user does not demonstrate any form of assent*

Browse-wrap licences add a further layer of complexity because the user does not demonstrate any form of assent. The validity of such licences was considered in the Canadian case of *Canadian Real Estate Association ('CREA') v Sutton (Quebec) Real Estate Services.*[66] CREA sought an interim injunction to prevent Sutton scraping its site. CREA argued that Sutton had ignored its terms of use and circumvented the technological measures that it had put in place to prevent data mining of its property information. The court found in CREA's favour, granting the injunction on the basis that Sutton was familiar with website terms of use and actually used similar provisions on his own site.

As in *CREA*, the US courts considered website terms of use and data scraping in *Register.com Inc v Verio Inc.*[67,68] There the court found that Verio's continued use—data scraping—amounted to assent to be bound to Register.com's terms of use. A further factor in the *Register.com* case was that Register.com expressly stated in its terms of use that data scraping was prohibited.

*Pro-CD* and *Register.com* indicate the extent to which that US judges have been prepared to shape the law in order to meet the demands of the information age. Indeed, the overriding message from *Register.com* case is that internet users should be alive to the fact that terms of use may be binding upon them, notwithstanding that they have not read or expressly assented to them.

[61] *Parker v South Eastern Railway* [1877] 2 CPD 416.

[62] *L'Estrange v Graucob* [1934] 2 K.B. 394.

[63] 86 F. 3d 1447 (7th Cir 1996).

[64] 104 F. Supp. 2d 1332 (D. Kan. 2000).

[65] 150 F. Supp. 2d 585 (S.D.N.Y.2001).

[66] Available to download at canlii.org.

[67] 356 F. 3d 393 (2nd Cir. 2004). The facts relating to this case are discussed further below, Trespass to Chattels.

[68] The facts relating to this case are discussed further below at 'Trespass to Chattels'.

In EU member states, the formation of contracts online is also subject to the Electronic Commerce Directive[69] (the 'E-Directive'). The E-Directive obliges all member states to ensure that

(a) their legal systems allow contracts to be concluded by electronic means and

(b) the legal requirements applicable to the contractual process neither create obstacles for the use of e-contracts nor result in such contracts being deprived of legal effect by virtue of their having been made by electronic means.[70]

It may be argued, therefore, that English courts are further obliged to follow the line taken in *Pro-CD* and *Register.com* in relation to shrink-wrap and browse-wrap licences, respectively.

The E-Directive also sets out at Article 11 the manner by which a contract is concluded online:

> in cases where a recipient, in accepting a service provider's offer, is required to give his consent through technological means . . . the contract is concluded when the recipient of the service has received from the service provider, electronically, an acknowledgement of the recipient's acceptance.

It may be stretching the construction of Article 11, but a website owner who is the victim of data scraping may attempt to argue that the data scraper had concluded a contract with the website owner: if the website owner has made a unilateral offer to provide certain services subject to its terms of use, the data scraper in sending an electronic message to the website owner signifies its assent to those terms of use, the contract being concluded when the website owner transmitted information to the data scraper. The validity of this argument remains untested.

English law has never suffered from an inability to recognize contracts formed where the parties do not actually meet, or where acceptance is not actually communicated. Also, given that most website owners now post terms of use, it is difficult for a user to argue that, notwithstanding common practice on the internet, he is unaware that his access to the website is governed by such terms. While requiring the user to click a box would leave the issue beyond doubt, many website owners only resort to this where they are selling goods or services to the user.

Website owners who wish to maximize the chances of being able to enforce their terms of use against 'regular' users and data scrapers alike should

(i) display a prominent notice at the outset of any online transaction on the homepage where terms are located;

(ii) stipulate precisely what conduct constitutes an acceptance of the terms of use;

(iii) ensure the terms of use are in plain language that is non-legal in nature and that they highlight any onerous provisions;

(iv) warn users that the terms of use are subject to change;[71] and

(v) implement the robot exclusion protocol.[72]

Experience suggests that website owners are more likely to bring a claim for infringement of IP rights, breach of contract being merely a subsidiary claim. This is probably because of the limited remedies available for breach of contract, the need to demonstrate loss, and the common law duty to mitigate against that loss. If an action for infringement of IP rights is successful, the scraped party can ask the court to award damages taking into account its lost profits and any unfair profits made by the infringer and possibly the moral prejudice caused to the scraped party by the infringement.[73]

## Trespass

The internet is a creature of the late 20th century. However, common law principles stretching back hundreds of years have a role to play in combating cybercrime—at least in the USA. US litigants have revived the law of trespass to chattels which, despite its age, is proving surprisingly responsive to countering data scrapers. Trespass to chattels arises out of the unauthorized use, dispossession, or interference with the tangible property of another. Although physical contact is required to constitute a trespass, the definition of physical contact has been greatly relaxed in the USA when compared to the English law,[74] where electronic signals sent down a telephone line between exchanges have been found to constitute trespass.[75]

In *eBay v Bidders Edge*,[76] the court granted eBay an injunction to stop Bidders Edge using spiders to pull

[69] Directive 2000/31/EC. This is enshrined in English law as the Electronic Commerce (EC Directive) Regulations 2002.

[70] Directive 2000/31/EC Art 9(i).

[71] Points (i) to (iv) are based on comments in HA Deveci, *Consent in online contracts: old wine in new bottles* (2007) CTLR 13(8), 223–231.

[72] Further information available from http://www.robotstxt.org/ (accessed 28 July 2008).

[73] Once Directive 2004/48 on the enforcement of IP rights is implemented. See Art 13(1).

[74] Under English law, the claimant must establish that the defendant has caused physical damage, no matter how slight, to the claimant's goods as a result of the defendant's direct and intentional interference: *Fouldes v Willoughby* (1841) 8 M & W 540 at 549 per Alderson B.

[75] *Thrift-Tel v Bezenek* 54 Cal Rptr 2d 468 (1996).

[76] 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

data from its site. Bidders Edge harvested data from a number of auction sites and aggregated the data allowing its users to observe the content of all of the various sites at once. eBay had initially consented to Bidders Edge carrying out limited searches on its site (for soft toys) using spiders. Bidders Edge, however, wished to expand its searches and approached eBay for consent to include all products which appeared on eBay. When negotiations between the parties broke down, Bidders Edge began searching eBay without consent. eBay claimed that the excessive searching carried out by Bidders Edge was causing it irreparable harm in that Bidders Edge's scraping resulted in (i) lost capacity of eBay's computer systems; (ii) damage to eBay's reputation and goodwill; (iii) dilution of eBay's mark; and (iv) unjust enrichment of Bidders Edge.[77]

Although disputing the effect of its searches on eBay's capacity, Bidders Edge admitted that it was carrying out something in the region of 80,000–100,000 searches a day, which represented about 1.2% of the traffic on eBay's databases.

The court agreed that the signals sent by Bidders Edge to retrieve information from eBay's website were sufficient to support a trespass to chattels and granted a preliminary injunction. Unfortunately, there was no full trial on the issues as Bidders Edge settled the claim and refocused its business activities away from eBay.

Trespass to chattels was also argued in *American Airlines Inc v Farechase Inc*,[78] where American Airlines argued that the Farechase's use of spider programs on the American Airlines website deprived it of some of its capacity which amounted to a trespass. It also maintained that, as the customer would deep-link direct to the booking page instead of first navigating through American Airlines' preliminary pages, the spidering of its pricing information deprived it of an opportunity to establish goodwill by building a relationship with its customers. A temporary injunction was granted, restraining Farechase from scraping its site.

This ancient tort may have a place in the USA, but it is unlikely that the English tort of trespass to goods would offer a UK-based website owner a chance of redress unless a claimant website owner could demonstrate that scraping resulted in demonstrable physical damage to its property. A website owner is far more likely to succeed by bringing a claim for copyright infringement, where there is no requirement to show physical damage.

## Computer misuse

The UK Parliament hastily enacted the Computer Misuse Act 1990 ('CMA') in the wake of *R v Gold and Schifreen*,[79] where the defendants successfully appealed against convictions for computer hacking under the Forgery and Counterfeiting Act 1981. The House of Lords considered that the Crown had twisted the meaning of the then-current legislation to such an extent that it could not uphold the convictions.

The principal offences under the CMA are

(a) unauthorized access to computer material;[80]

(b) unauthorized access with intent to commit or facilitate commission of further offences;[81] and

(c) unauthorized modification of computer material.[82]

Parliament did not define 'computer', 'program', or 'data', which allows a degree of flexibility in the application of the CMA.

The unauthorized access offence may be most pertinent when considering data scraping, applying where a person causes a computer to perform any function with intent to secure access to any program or data held in any computer, if the access he intends to secure is unauthorized and he is aware of this.

Arguably, a data scraper satisfies the three limbs of the unauthorized access offence. The difficulty for the scraped party is that it would either have to convince the Crown Prosecution Service of the merits of such a prosecution or bring a private prosecution itself. A further disadvantage concerns the Act's territorial limitations: since the offending data scraper may be located anywhere in the world, the likelihood of successfully identifying a determined data scraper and managing to prosecute it within the UK is small.

While the CMA has not been directly applied against data scrapers, it was linked to the withdrawal of Citibank's first venture into online account aggregation services in the UK.[83] Citibank launched its 'My Accounts' service in 2001, which allowed its users to aggregate all their online accounts on to one webpage irrespective of the bank that held the account. Citibank's system relied on users providing it with their account details and

[77] See eBay's motion for a preliminary junction cited in 100 F. Supp. 2d 1058.

[78] Cause No. 067-194022-02, 67th District Court, Tarrant County Texas, 12 February 2003.

[79] [1988] 2 ALL ER 186.

[80] CMA s 1.

[81] CMA s 2.

[82] CMA s 3.

[83] http://www.computerweekly.com/Articles/2002/05/31/187519/citibank-overhauls-online-aggregation.htm (accessed 23 August 2008).

passwords. Citibank would then use 'bots' to access the various accounts as though it was the user. Other banks, concerned that a competitor such as Citibank was using such means to make unauthorized access to accounts they held, alleged that Citibank was committing an offence under the CMA. Citibank relaunched the service in 2002 and stayed within the law by getting its customers to install a 'data safe' on their computers which contained the relevant passwords which, when activated by the customer, would harvest the relevant data from their various accounts.

The US equivalent of the CMA—the Computer Fraud and Abuse Act 1986[84] ('CFAA')—has been applied against data scrapers. The CFAA makes it an offence knowingly to use a program or access a computer without authorization and, as a result, cause more than $5,000 of damage. The CFAA differs from the CMA in that it also allows for those suffering harm to bring a civil claim. Thus, website owners whose data are scraped may bring a claim provided that the standards of the criminal test are met, and the claim is brought within 2 years. Where a website owner brings such a claim, it must demonstrate that harvesting of data from the website is not authorized.

Examples of successful applications of the CFAA against data scrapers include *Register.com, Inc v Verio, Inc*[85] and *America Online, Inc v LCGM, Inc.*[86] In *Register.com*, Register.com sought and obtained an injunction to prevent Verio[87] accessing its publicly available Whois database.[88] Likewise, AOL[89] obtained summary judgment against LCGM, a notorious spammer. LCGM scraped AOL's files for email addresses to spam. AOL was successful because it was able to demonstrate that its terms of use specifically prohibited users from using spiders to obtain email addresses for the purpose of spamming.

## Data protection

As well as disseminating data and information, websites collect information from their users. This may include basic customer data required to fulfil an order or, in the case of social networking sites, sensitive personal information. One can garner an impression as to the amount of personal data available online by reference to the dramatic growth in Facebook use. Between June 2006 and June 2008, the number of visitors accessing Facebook grew from 14 million to 132 million, there being 581 million total social networking users.[90]

Given the volume of personal data potentially available to a data scraper, it is pertinent to consider the legal position of website owners and data scrapers against the backdrop of the Data Protection Directive[91] and its UK implementation, the Data Protection Act 1998 ('DPA').

The principal objectives of the DPA are (i) to ensure that those who process information concerning individuals do so lawfully (including website owners collecting personal data) and (ii) to provide individual website users with enforceable rights to protect their personal data.

The processing of information or data includes retrieval, consultation, or use of personal data.[92] The DPA obliges those processing personal data to take appropriate technical and organizational measures against unauthorized or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.[93] Further, section 55 of the DPA makes it unlawful to obtain personal data without the consent of the data controller. It is therefore possible, if personal data is scraped from a website, that both the owner and the scraper may face action from the Information Commissioner's Office ('ICO'), which has legal powers to ensure compliance with the requirements of the DPA.[94]

Given the huge growth of the likes of Facebook,[95] it is not surprising that social networking sites have attracted the attention of the ICO, which has even launched a 'youth friendly' site to advise teenagers of data protection issues and social networking.[96]

Social networks are entitled to process personal information since the user consents to such processing.[97] For example, Facebook's own terms of use,[98]

[84] 18 USC 1030.

[85] 126 F. Supp. 2d 238 (S.D.N.Y. 2000).

[86] 46 F. Supp.2d 444, Civ. Act. No. 98-102-A (E.D. Va., 10 November 1998).

[87] Register.com's claim also included breach of contract, passing off (under the Lanham Act) and trespass to chattels.

[88] The Whois database recorded details of domain names registered by Register.com. Verio wanted to use this information to market its own services.

[89] AOL additionally asserted violation of the Lanham Act, trade mark dilution, violation of the Virginia Computer Crimes Act, and trespass to chattels.

[90] http://www.associatedcontent.com/article/333091/myspace_facebook_bebo_increase_in_popularity.html (accessed 23 June 2008), http://www.insidefacebook.com/2008/07/27/intriguing-trends-in-social-networking-growth-during-1h-2008/, http://www.insidefacebook.com/2008/07/29/tracking-facebooks-2008-international-growth-by-country/

[91] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[92] DPA s 1 (1) (b).

[93] DPA, Schedule 1 (7th data protection principle).

[94] DPA s 40.

[95] http://www.facebook.com

[96] http://www.ico.gov.uk/Youth.aspx (accessed 23 August 2008).

[97] DPA s 7.

[98] http://www.facebook.com/terms.php (accessed 23 August 2008).

which a prospective user must accept when activating a Facebook account, provide that the user consents to Facebook and third party developers of applications having access to personal information given by the user. The user may, however, alter privacy settings to restrict use of personal information. A difficulty that Facebook has encountered, whilst not data scraping in a traditional sense, is the use of a Google application (Friend Connect) which allows a user to export information from Facebook accounts of friends to third party websites. The ICO has stated that there is an urgent need to update European data protection legislation to tackle these new technologies.[99]

News organizations have also come under attack for extracting personal data from social networking sites.[100] While these actions do not constitute data scraping as such, they show that the unauthorized use of personal information from websites may put both the owner and the scraper in breach of the DPA. The DPA provides exceptions for the processing of personal data for special purposes.[101] Since those special purposes are journalism, literature, and art,[102] news organizations may have a defence to such use of personal data. The same would not be true of a data scraper.

Data protection issues have also been raised in relation to online aggregation services.[103] It has been suggested[104] that such services may place an aggregator in breach of the DPA in that

- the aggregator's notification to the ICO may not include such processing;
- the aggregator may not have consent to use the data for commercial purposes; and
- the aggregator may not have taken sufficient precautions to protect personal data from data scrapers.

## Where next for data scraping?

From the scraper's point of view, scraping software can be used to create attractive content at little cost and allows information to be gathered on individuals.[105] This article has shown that content owners have a variety of legal recourses open to them to remedy scraping after it has occurred. Perhaps the most sensible route is to adopt measures that prevent scraping in the first place.

Yell.com has a database of more than two million classified UK business listings and is alive to the potential costs of data scraping. In 2006, Yell.com introduced a security system called ASSASIN to combat the growing threat of data privacy. ASSASIN monitors Yell.com's database and detects and blocks systematic searches and downloads from the database. Yell.com announced that it does not license the data for commercial or users' personal gain (for example, the downloading of data that are then used for sales prospecting or telemarketing).[106] Also, Yell.com described actions by third parties such as screen scraping as an infringement of the company's IP rights, voicing concern that data scraping could mean that its advertisers would receive unsolicited calls.[107]

Additionally, while it has been successful in the courts, Ryanair has announced that it will cancel tickets booked through aggregator or scraper sites.[108] This is a bold step and one which may damage its reputation with its ultimate customer—the individuals who have paid for the flights. Bravofly has denied that it has stopped its activities following the filing of the law suit.[109] Ryanair is looking at technological means of preventing the data scraping from happening in the first place.

The battle between the content owners and the data scrapers is now raging. While it is clear that the scrapers have won the first battle, the content owners have fought back with victories for Register.com, easyJet, and Ryanair. While content owners who do not adopt security measures can rely upon the courts to protect their rights, it seems that the best method of protection is to introduce security measures to prevent the scrapers copying the content in the first place.

*doi:10.1093/jiplp/jpn232*
*Advance Access Publication 00 Month 2008*

[99] http://www.ico.gov.uk/upload/documents/pressreleases/2008/ico_leads_debate_070708.pdf (accessed 23 August 2008).

[100] http://news.bbc.co.uk/1/hi/technology/7271348.stm (accessed 23 August 2008).

[101] DPA s 32.

[102] DPA s 3.

[103] J Chuah, 'Internet banking services—questioning the current response to account aggregation' *JIBL* 2002, 17(11), 309–315. See also discussion regarding Citibank's 'My Accounts' service above.

[104] Worthy and Graham, 'Account aggregation—avoiding the pitfalls in internet banking' (2002) 4 *Finance & Credit Law* 1.

[105] Financial Services Authority, 'The FSA's approach to the regulation of e-commerce' (June 2003) 9.7–9.14, http://www.fsa.gov.uk/pubs/discussion/dp6.pdf (Accessed 23 August 2008).

[106] Yell Group, Media Press Release, 6 June 2006.

[107] Brand Republic (*section 9(1), CDPA*) In depth News by David Murphy, 4 April 2007.

[108] http://www.ryanair.com/site/EN/news.php?yr=08&month=aug&story=reg-en-050808

[109] http://www.bravofly.com/content/en/pdf/11-08-08_Press_Release_Bravofly_EN.pdf