Data Scraping

Expert knowledge means success

Contents

Note: This publication has not been updated since it was last published. Some of the hyperlinks may have changed and may need updating. In addition, some of the information in this publication may be out of date.

Introduction

"Scraping" is an umbrella term used for a number of separate but related types of activity. In short, it covers the automatic extraction of data from a computer program by another computer program.

Scraping is a technique often used without the knowledge or permission of the data owner and includes:

- Screen scraping - the extraction of relevant data from the code that generates the display output of a program (i.e. the screen). Screen scraping programs are developed to extract the core text on a web page, removing irrelevant style information and deliver the extracted data to the scraping party;
- Web scraping - involves extracting all the data relating to the underlying structure of the script used to create that website and not just the data displayed on screen;
- Web-harvesting - the name given to programs that extract data by following links from one website to another. There are several names given to these programs are known as "bots", "webbots", "crawlers", "harvesters" or "spiders".

In addition to scraping, content may be acquired by "deep linking". A "deep link" effectively captures or uplinks data from one website to the other by-passing the homepage of the originating website, often making it appear to be part of the linking website.

Commercial advantages of scraping

Scraping data may secure many commercial advantages for a scraper including:

- content creation at little to no cost;
- attractive content which will in turn drive traffic to the site;
- increased traffic which may result in increased advertising revenues.

For the website owner whose site is scraped, scraping may cause a number of undesirable consequences including:

- system overload;
- loss of advertisement revenue;
- loss of control of the information content;
- devaluation of content (particularly if that content was from a premium service).

Legal issues

Although there is little UK case law on scraping, the legal issues are likely to include:

Copyright infringement

Scraping is, in effect, copying and this may lead to a claim for copyright infringement. Whether such a claim has any merits will depend on the particular circumstances. For example, not all scraped data qualifies for copyright protection.

Copyright infringement has also been raised in relation to deep linking. In 2006, a US District Court held that an unauthorised link from one website to the live webcasts of another was likely to be copyright infringement. The lack of UK case law may evidence the fact that infringing linkers recognise that they have no defence to a claim and settle at an early stage.

Database right infringement

The database right is distinct from copyright and has become increasingly important as the internet/information services have developed. The right arises in a database where there has been a substantial investment in obtaining, verifying or presenting the contents of a database. A database right is infringed when all or a substantial part of a database is extracted or re-utilised without the consent of the owner.

Database rights' infringement may be particularly applicable to scraping directory or listings type information from third party websites - such information may not qualify for copyright protection but may qualify for database right protection if the owner has incurred expenditure in developing and maintaining the database.

Also note, "substantial" relates to quantity and/or quality. Further, the repeated extraction/re-utilisation of insubstantial parts of a database may in fact constitute a substantial

Data scraping software

Many businesses are finding that their customers are using the internet more than ever to check prices. Customers often say that the product they want is available cheaper online. To overcome this, businesses are turning to data scraping software to check competitors' websites and prices.

part. Thus, use of a web-harvester to repeatedly interrogate the same database may infringe database rights.

Data protection legislation

The control and processing of personal data – that is, information about a living person – is regulated by the Data Protection Act. Acquiring personal data as a result of data-scraping may breach this Act as the data subject is unlikely to have consented to this type of data processing. Moreover, last year the FSA issued a large fine against Nationwide losing customer data.

Breach of contract

In order to protect data, owners of websites have attempted to enforce website terms of use which prohibit scraping. The enforceability of such terms of use is problematic. For example, were the terms prominently displayed? Did the viewer (ie the scraper) agree to the terms before accessing the site? The scraper may argue that it had no notice of such terms.

Conversely, a website owner might be able to argue that it is common practice to adopt terms of use and that web-users should expect there to be some applicable terms. Adding a link to these terms from the homepage might be sufficient to put the web-user on notice. It is possible, therefore, that scraping might amount to a breach of contract.

Computer misuse

The Computer Misuse Act makes it a criminal offence to access a computer program or data without authorisation. The offence is wide-ranging in nature and includes "hacking" or even the simple use of someone else's password to access a computer. It is possible that the CMA will catch data scraping as the website owner will not authorise the type of access made by the scraper. In order to be guilty of an offence, the scraper must be aware that such access to the computer program or data is unauthorised. The scraper might argue that, by making the relevant data available to the public via its website, the owner has granted a licence (authorised) for all to access the data. However, this argument is not likely to be successful as any such implied licence would not be deemed to extend to scraping.

An example of the impact of the CMA can be seen in relation to bank account aggregation services (a service which allows a bank customer to receive information about its various bank accounts). In September 2001, Citibank were the first bank to launch such a service in the UK. Citibank fell foul of the CMA due to the fact that its service automatically updated the users' information on a daily basis without asking permission from the banks from whose computers they were "scraping" data. This problem was rectified by removing the automatic update facility.

Outcome of a successful legal action

Unauthorised scraping could give rise to a number of remedies including:

- an injunction preventing further use of the infringed material;
- payment of damages or an account of profits derived from the scraped data;
- fines under data protection or computer misuse legislation;
- imprisonment under computer misuse legislation

Conclusion

Data scraping offers many obvious commercial advantages but, unless done legally or with the agreement of the owners of the websites, it is likely to be unlawful. It is worth noting that websites, such as copyscape.com, allow content comparison so that a website owner may determine if material has been copied to another site.

The scraper must consider the impact that any publicity surrounding its (unlawful) use of data-scraping may have on its reputation or other trading relationships. Publicity, of course, is a double-edged sword – what may appear to be a negative may in fact be a positive if it drives more traffic to the scraper's website. However, if a scraper faces a claim for infringement copyright or database right, it should be prepared to settle claims at an early stage.

It may also be important for the scraper to consider what it may be able to offer the other side in such circumstances – e.g. reciprocal data sharing.

Scraping data from websites is on the rise. If this scraping is done and the data re-used without the owner's consent, this could leave the scraper on the wrong end of the law.

Definition of "Screen Scraping"

"Screen scraping is a technique in which a computer program extracts data from the display output of another program. The program doing the scraping is called a screen scraper. The key element that distinguishes screen scraping from regular parsing is that the output being scraped was intended for final display to a human user, rather than as input to another program, and is therefore usually neither documented nor structured for convenient parsing. Screen scraping often involves ignoring binary data (usually images or multimedia data) and formatting elements that obscure the essential, desired text data. Optical character recognition software is a kind of visual scraper.

There are a number of synonyms for screen scraping, including: Data scraping, data extraction, web scraping, page scraping, web page wrapping and HTML scraping (the last four being specific to scraping web pages)."

Source: Wikipedia

Further Information

This guide is for general interest - it is always essential to take advice on specific issues. We believe that the facts are correct as at the date of publication, but there may be certain errors and omissions for which we cannot be responsible.

Acknowledgement

[1] © We acknowledge that the body text to this publication is copyright to DMH Stallard

DMH Stallard LLP is one of the country's leading law firms with 56 partners and around 300 staff. With offices in London, Gatwick and Brighton they act for clients nationally and internationally. DMH Stallard is recognised in legal directories and journals as a Top 100 Law Firm with a great track record. They can be contacted at:

- London: +44 (0) 20 7822 1500
- Gatwick: +44 (0) 1293 605000
- Brighton: +44 (0) 1273 329833

For legal advice on the subject covered by this publication, please contact: Frank Jennings (Tel: +44 (0) 1293 605018 (DDI) or e-mail: frank.jennings@dmhstallard.com) or John Yates (T: +44 (0) 1293 605591 (DDI) or e-mail: John.Yates@dmhstallard.com).

Important Notice

Publication issued or updated on: 27 January 2012

Ref: 755