

Screen Scraping Strategies A Management Guide

June 2004

Scraping of property listings and member information from web pages is an increasing problem in the real estate industry. Implementing an effective strategy to prevent this should start with gathering interdisciplinary input including policy, legal and technical perspectives.

Effective strategies for combating this problem includes both reactive and proactive elements. Reactive tactics are employed after you think your site is being scraped and require a mix of legal and research resources. Proactive tactics are used to prevent scraping abuse and require investment and technical resources. Reactive measures taken after you already suspect you have been compromised are expensive and time consuming. If you take proactive measures, they will allow you to avoid some, but not all, reactive situations.

Important trade-offs between consumer utility and effectiveness have to be made when implementing proactive measures. This trade-off also applies to the installation and maintenance of your website. Although not practical, the most secure web site is actually one that does not display any data. Technical approaches that provide a balanced trade-off between utility and defeating screen scrapers are rendering and CAPTCHA.

How Scrapers Work

Screen scrapers are simplistic. Pirates typically employ simple exploits because less complex tactics are more reliable complex ones. They are not concerned about completeness or accuracy of the information that is gathered. Their goal is to obtain the data that in turn will give their consumer the impression that they have legitimate access to information. The consumer is unaware of the quality of the data.

If you have ever highlighted text or images from a web page, and saved them onto your hard drive, you have performed a form of screen scraping. Pirates do not use browsers though. The term for the tool they use is “bot”, short for robot.

Bots pretend to be browsers, and your server can't tell the difference. Instead of rendering a web page though, bots extract the data and images and save them to the thief's hard drive. Technically, this process involves parsing the HTML supplied by the server and extracting the data and image elements. While searching for information on the Internet, bots only retain the data elements that they find, discarding other "markup" elements of HTML. This is an important operating characteristic that can be taken advantage of when designing proactive measures.

Search engines scan the Internet with similar technology. There may be a trade-off to consider because moving the search capability to a less prominent section of your website may have the unintended consequence of being missed by search engines like Yahoo and Google.

Proactive Measures

There are currently three common proactive measures that can be taken to prevent screen scraping: *Posting*, *Cleanup* and *Investment*. Each of these measures by themselves have varying degrees of effectiveness but when used together can make the pirate's task much more difficult.

Posting advises those accessing your site that screen scraping is not allowed. You should prominently display your policy concerning site usage. Although this would not stop scraping in a technical sense, it is an effective approach to stop those who might claim that they didn't know they weren't allowed to scrape your site. It is also a necessary step to take if you plan to execute reactive measures if you find yourself being scraped by demonstrating that notification was posted prior to the pirate beginning the theft.

Cleanup involves checking your current website for exposed e-mail addresses, passwords, account numbers and other authenticating information. Removing easy to get data makes screen scraping and social engineering more difficult.

During the cleanup process, you may consider "watermarking" your images. This involves using software to alter the visual part of the image to embed a pattern that is uniquely yours. The end result is barely noticeable to the consumer and once watermarked, your content can be identified on the Internet. You should be aware that most Digital Watermarks can be removed with consumer grade "painting" software.

Investment is the most expensive and involves implementing a technical barrier to screen scraping. It is the only way to directly defeat bots. There are four barrier types: *Limiting*, *DRM*, *Rendering* and *CAPTCHA*. Each of these approaches have their own degrees of effectiveness, ease of implementation and consumer impact.

Limiting is a popular defense to screen scraping used today. Servers are changed to only show a small number of results at one time or limit the number of results that are returned from queries. Unfortunately, bots can easily work around *Limiting* approaches. This approach ends up limiting access to legitimate users and does not deter those who screen scrape. It can also give you false sense of security.

DRM is short for Digital Rights Management and is currently the barrier favored by the Music and Video industries. The approach is characterized by downloaded “plug-ins” and “readers”. Summarizing the *DRM* approach is simple, the downloaded program disables specific functionality of the connected computer based on rules provided by the information supplier.

You should be aware that the potential for consumer backlash is highest with the *DRM* approach. Software that you purchase, or write, will reside on the consumer's computer. Many who argue against *DRM* liken the downloaded programs to “spyware”.

Rendering is a new approach that is being used to defeat bots and is favored by the on-line financial institutions, coupon and ticketing industries. This approaches the problem creatively by differentiating data from information. Consumers typically use the Internet to gather information, not data. Most websites generate data filled HTML that the consumer's browser then turns into information in a visual form.

Bots are designed to gather web pages and strip out the data embedded within the HTML. *Rendering* generates an image that contains the combined data and image. When servers deliver content that is was already rendered, bots can not simply strip the data from the HTML.

CAPTCHA is short for Completely Automated Public Turing test to tell Computers and Humans Apart and approaches the problem in a unique way. It identifies the party trying to access your site as a human or a computer program. It does this by generating questions that only a human can answer correctly. Computers still lag humans in the area of image and word recognition. *CAPTCHA* displays distorted images of a word and challenges the party to correctly enter the word. You should be aware that there have been academic algorithms published that can defeat some forms of *CAPTCHA*. As with rendering, you will find *CAPTCHA* being used by financial and ticket industries.

Reactive Measures

If you believe that your website is being scraped, there are two reactive tactics your organization can execute. Both approaches require you to identify the individual and given the anonymous nature of the Internet, this can be a difficult task.

Identity comes in two flavors, *Physical Identity* and *Internet Identity*. You will need to gather one of these two forms to implement reactive measures. *Physical Identity* is based on a name, address or phone number. *Internet Identity* is characterized by a URL, e-mail address or IP address. If you know one form of identity, it is difficult to get the other.

Periodic policing of the Internet can help to identify organizations who you believe have a copy of your data. Either you can perform this policing, or your Member Service group can be trained to field phone calls and e-mails from the members who report the situation. The more eyes that watch the Internet, the better.

The information you are looking for here is the business entity. Information about the business entity can lead you to the *Physical Identity*. Once identified, the best measure is to send a registered letter to the party informing them of the situation. You should have your legal staff involved in wording this letter.

Internet Identity is difficult to trace to a *Physical Identity* but can still be useful by itself. The following steps can help you take action against the offender and involve contacting someone who can actually take action. If you know a URL or an e-mail address, the domain name can be determined. Once you have the domain name you can contact a registrar (like register.com) to look up the either the owner of the domain or the technical contact.

Once you know the contact for the domain, you should send notification. Again your legal staff should determine the wording of this e-mail. Notification is the most effective first reaction and can set the ground work for more intensive legal measures. Pirates often react to the notification so you should develop a formal policy that is easy to execute.

If notification fails, more drastic actions can be taken. Your firewall can be set to block access for specific IP Addresses. If all you have is the domain name, the Internet registrar can tell you what IP Addresses are used by the domain. You should keep your notifications on file to help identify repeat offenders.

Conclusion

There is no single solution that can be used to prevent the theft of images and data from your website. The best defense is to develop a strategy that combines both Reactive and Proactive measures. These measures are most effective when a mix of legal and technical tactics are used.

One critical factor to the success of your strategy will be determining the true identity of the offending party. Another factor is that you must plan for the current approaches to screen scraping to evolve. You must review your

approaches on a routine and on-going basis.

