

# Sins Of The Internet: Email Spiders

By Richard Lowe

Sins Of The Internet: Email Spiders

Richard Lowe  
articles@internet-tips.net

Internet Tips And Secrets <http://www.internet-tips.net>

Copyright (C) Richard Lowe Jr. and Claudia Arevalo-Lowe, 1999-2001.  
Permission is granted to reprint the following article as long as no changes are made and the byline, copyright information, and the resource box is included. Please let me know if you use this article by sending an email to <mailto:articles@internet-tips.net>

Article Title: Sins Of The Internet: Email Spiders  
Author: Richard Lowe, Jr.  
Contact Author: <mailto:articles@internet-tips.net>  
Publishing Guidelines: May be freely published w/bylines  
Web Address: <http://www.internet-tips.net>  
Autoresponder Address: <mailto:article-180@internet-tips.net>

Warning: this article is not for the squeamish. It contains graphic descriptions of one of the biggest evils on the internet. If you can face down this evil you can reduce your load of spam by several times. Hold onto your seats and try and keep down your lunch - you are about to learn one of the secrets of how ruthless, unethical and, well, downright evil spammers steal your email address - and what you can do about it.

If you have access to your web site's log files, you will quickly find an interesting phenomenon. Your site is being visited a lot more often than you think it is. In fact, if you look closely you may be shocked to find that your HTML files are actually being used to harm you and others. In fact, you may be seeing the footprints left by some of the tools used by unscrupulous spammers to steal your email addresses.

Oh wait, let me back up a bit and explain a few things. Each time you visit a web site a record is kept of every page, graphic, sound file, video or anything else that you access (look at or download). This record is called a log file. Each line within the log file is one "hit" (other things are recorded also, but that is not important to this discussion). A "hit" is getting one "thing" from a web site. A "thing" can be an image, an HTML page, a video, a sound file or anything else. In fact, generally when you look at one HTML page you are actually "hitting" the web site many times, once for each file on the page.

Each of these lines within the log file records a number of pieces of information so that webmasters can later see what happened (don't worry, they are not generally interested in individuals - they want to know things like how many people are using Internet Explorer verses Netscape). One critical

piece of information is called the "user agent". Generally this contains the browser name (Internet Explorer for example) or spider name (googlebot, for example, is the spider for the Google search engine).

Examine these user agent fields and you will find out many interesting facts. You will see that your site is being visited a lot more often than you would think by lots of things with strange names:

- Googlebot
- Slurp (used hundreds of search engines including Hotbot)
- Scooter (Altavista robot)
- Lycos Spider (used by the Lycos search engine)
- and many others as well.

Most of these are innocent 'bots, used by the major search engines to keep their indexes up to date. These robots are very important, for they keep your pages listed so you will get traffic. Occasionally they have other uses, including checking your pages for changes, saving your pages for offline browsing and various statistical functions.

You will also find some other names buried in your log files. These go by names such as EmailSiphon and Cherry Picker. These robots are malignant and are used by spammers to harvest email addresses. What they do is scan every single page in your web site, as fast as they can, looking for email addresses. Specifically, they are usually looking for "mailto:" type links.

Many websites have these kind of links. They are convenient, simple and create a great way for visitors to send an email to someone. In fact, it's hard to find a website which does not have email addresses embedded somewhere within the site.

In addition, people often leave their email addresses in guestbooks, message boards and other online communities which translate to web pages. Spam harvesters love these types of pages, as they can get dozens, hundreds or even thousands of different, valid and usable email addresses quickly and easily.

How do email harvesters work? Well, some scum spammer will install one of these programs on his system. He will tell it to begin scanning, which it will do rapidly and efficiently. In fact, these generally scan a web site so quickly that the server cannot do anything in the meantime (most "good" spiders, on the other hand, limit their visits to one per second, minute or even hour in order to allow other people and spiders to use the site while it is being scanned).

One of the more popular email harvester programs is called EmailSiphon (a product known as Sonic). The web site which promotes this garbage has the following to say:

"First of its kind on the market, Sonic helps you extract highly targeted email addresses from World Wide Web pages. Earthonline Internet marketing expertise has enabled us to program a powerful, yet sensible product that allows for proven focused lead harvesting. Therefore, Sonic with its search engine ability and single domain capability is only second in World Wide Web extraction to Earthonline Nitro."

Obviously these scumbags think they are doing a great service to the world by providing the opportunity to scan thousands of sites per day for email addresses.

Okay, so what can you do?

Ask them politely - With most "good" spiders, this is very easy to do. You simply create a robots.txt

file or use the robots metatag (if you don't know what those are, don't worry about it). Unfortunately, email harvesters are written by and used by scum, so they typically ignore polite requests.

Block them - On some web servers this is possible using special commands in a file called htaccess (again, don't worry about it if you don't know what that is), but only with those robots that clearly identify themselves. For those that don't tell you who they are (and some of them do not), then you cannot block them. In addition, the web host has to be specially set up to allow you to do this - and most, in my experience, are not.

Confuse them - Some webmasters create page after page of fake email addresses. These pages are not intended for people or good spiders (the robots metatag is used to keep the good one's out) but rather are made attractive to email harvesters. The theory is simple - the harvesters will not be able to resist the temptation (they are not very bright, as programs go) and will scan these pages. They will grab dozens, hundreds and then thousands of fake addresses, thus wasting the spammers time and possibly causing their programs to crash.

Does this work? Sure - occasionally, but it also does not prevent the spammers from getting your other email addresses, and it still chews up resources (web servers and bandwidth) sending useless messages all over the internet.

Cloak your email addresses - One thing you can do that is fairly effective is to make your email addresses look like something else. Some people create a graphic image with the email address in it (not a great solution as it means the email address must be retyped by your visitors). Others use JavaScript to make the email address look like code. These solutions work (usually), but they make it difficult to maintain your site and often make it more difficult for your visitors. In addition, presumably the spam harvesters will eventually catch on and make their programs smarter.

Strip your site of email addresses - The only solution that works for the present time is to remove all email addresses from all of your web pages. If you need to get your visitors to send you information, then use a form (these cannot be harvested by spammers as long as the email address is not part of the form itself - Bravenet is a good service to use for this purpose). If you don't put your email addresses directly on your site, then the spammers cannot get it using their harvesters.

So there you have it. I hope this is of use to you in fighting this internet evil known as email harvesting.

NOTE: The following information must be included if you reprint this article:

-----  
Richard Lowe Jr. is the webmaster of Internet Tips And Secrets. This website includes over 1,000 free articles to improve your internet profits, enjoyment and knowledge.

Web Site Address: <http://www.internet-tips.net>

Weekly newsletter: <http://www.internet-tips.net/joinlist.htm>

Daily Tips: <mailto:internet-tips@GetResponse.com>

Claudia Arevalo-Lowe is the webmistress of Internet Tips And Secrets and Surviving Asthma. Visit her site at <http://survivingasthma.com>

List of articles available for reprint: <mailto:article-list@internet-tips.net>

[Get-Articles.com](http://Get-Articles.com) : 1000's of reprintable business and internet marketing-related articles.

[Submit your article for reprint.](#)