No-Email-Collection Flag

Matthew Prince, Arthur M. Keller & Ben Dahl[†]

Unspam, LLC, Chicago IL 60607 USA

Abstract. One source major of email addresses for spammers involves "harvesting" them from websites. This paper describes a proposal to allow a website owner to make illegal such automated extraction of email addresses. We propose a "no-email-collection" standard for websites and a mechanism to track spammers violating the standard. There are four parts to this proposal. The first part is to create special restrictions for non-human visitors as part of the Terms and Conditions of Use page for the website. The second part is to add a Meta Tag specifying the "no-email-collection" flag referencing the Terms and Conditions of Use webpage to each web page of the website. The third part is to add "no-email-collection" flag referencing the Terms and Conditions of Use webpage to the robots.txt file. Finally, the fourth part of the proposal provides instructions on the creation of "honey pot" email addresses to track violators.

1 Introduction

In order to stop spam, we need to move earlier in the "spam cycle." Specifically, we need to stop spammers before they first get their hands on a list of email addresses. Spammers currently use a number of tools to scour websites and newsgroups to automatically gather email addresses. [FTC] While technological solutions for stopping these spiders, crawlers, harvesters, indexers, and other bots are needed, this problem may be a place where law can have a positive impact.

Up to this point, law has had a difficult time keeping up with the spam problem. Many believe there is no place for the law in the fight against spam. However, law has one advantage over technology in this fight: it can impose costs by creating risk. Since spammers face virtually no incremental costs for sending each additional message, their response to better filtering technologies has been to simply send in greater volume. Law can impose marginal costs on spammers. By setting a value on email messages, legal mechanisms can create a financial risk for abusive spammers.

Under the Federal CAN-SPAM Act [CSA], "harvesting" websites for email addresses is forbidden. Unfortunately, however, this area of the Federal law is untested and specifically limits an individual's right to sue. Courts often have a difficult time interpreting new Federal laws as matters of first impression because their rulings can have sweeping effects. As a result, courts tend to read Federal laws narrowly. This bodes poorly for any efforts to legally prohibit the harvesting of email addresses from websites. Moreover, websites are generally not configured with a mechanism to track harvesting of email addresses from their pages.

What courts are generally more willing to interpret and enforce are individual licensing contracts. Specifically, it is generally accepted that prominently posted Terms and Conditions of Use can be binding on the visitors to a website. "Prominently posted" ideally means that a link should be included near the top of every page and in a relatively large font. Because these Terms and Conditions of Use are seen as binding, they can effectively be used to

---

[†] Matthew Prince is the CEO of Unspam, LLC. He is also an Adjunct Professor at John Marshall Law School. Arthur M. Keller is a visiting associate professor of Computer Science at the University of California at Santa Cruz. Ben Dahl is the COO of Unspam, LLC and an attorney.

create a no-trespassing sign for certain visitors to your website. Specifically, we propose creating a “no-email-collection” flag that will apply to harvesting bots and other non-human visitors.

In order for this to be effective the flag must be accepted as an industry standard. As an analogy, a court will enforce a “No Solicitations” sign in your front yard in the United States if it’s written in English, but they are far less likely to accept it if it's written in Ancient Greek. In order for courts to accept this licensing term we must first get the technical community to accept it. Below we propose the beginnings of that this technical and legal standard.

2 The Proposal

We describe the parts of this proposal in more detail.

2.1 Terms and Conditions of Use Page

If you do not already have one, you should consider creating a “Terms and Conditions of Use” page for your site. In drafting a Terms and Conditions of Use, you should contact a lawyer that can address your particular website’s needs. In general, Terms and Conditions of Use pages contain regulations regarding the access to your site. Courts have generally recognized that visitors to a site are bound by these conditions, so long as a link to the terms and conditions are explicitly posted. For this reason a link to the Terms and Conditions of Use page should appear explicitly on each and every page of your site. Below is sample language for a Terms and Conditions of Use page that prohibits email collection and permits you to track abuse.

TERMS AND CONDITIONS OF USE

The website from which you accessed this agreement ("the Website") is provided to you subject to the following conditions. These terms are in addition to any other terms governing access to the Website. By visiting (in any manner) the Website you accept these terms and conditions (the "Terms of Service"). Please read them carefully.

Any Non-Human Visitors to the Website shall be considered agents of the individual(s) who control or author them. These individuals shall ultimately be responsible for the behavior of their Non-Human Visitor agents and are liable for violations of the Terms of Service.

SPECIAL LICENSE RESTRICTIONS FOR NON-HUMAN VISITORS

Special restrictions on a visitor's license to access the Website apply to Non-Human Visitors. Non-Human Visitors include, but are not limited to, Spiders and Spider Masters. "Spiders" include, but are not limited to, web spiders, bots, indexers, robots, crawlers, harvesters, or any other computer programs designed to access, read, compile or gather content from the Website automatically. Spider Masters include, but are limited to, any individual directing, authoring, running, hosting, controlling or otherwise utilizing Spiders to access, read, compile or gather content from the Website. Such Non-Human Visitors are restricted from taxing the resources of the Website beyond what would be typical of a human visitor.

In consideration for access to the Website, Non-Human Visitors agree to read and observe the restrictions as set forth in the robots.txt file included at the root level of the Website. Those restrictions on Non-Human Visitors set forth in the robots.txt file shall be considered a part of the Terms of Service. The robots.txt file specifies restrictions to the directories Non-Human Visitors may access. Non-Human Visitors accessing directories beyond what is allowed by the robots.txt file is recognized by the parties to this agreement as a breach of the Terms of Service by the offending Non-Human Visitor.

Furthermore, as specified by the "no-email-collection" flag in the header of every web page and the robots.txt file, email addresses on this site are considered proprietary intellectual property of the author of the Website. It is recognized that these email addresses are provided for human visitors alone, and have value in part because they are accessible only to said human visitors. You further acknowledge and agree by accessing the Website that each email address the Website contains has a value not less than US $50 derived from their relative secrecy. The compilation, storage, and potential distribution of these addresses by Non-Human Visitors substantially diminish the value of these addresses. Intentional collection, harvesting, gathering, or storing email addresses by Non-Human Visitors is recognized under this agreement as a violation of this agreement and expressly prohibited.

APPLICABLE LAW AND JURISDICTION

Each party agrees that any suit, action or proceeding brought by such party against the other in connection with or arising from the Terms of Service ("Judicial Action") shall be governed by the law of the state of residence of the registered Administrative Contact (the "Admin State") for the Website as such laws are applied to agreements between Admin State residents entered into and performed entirely within the Admin State. The visitor to the Website consents to the jurisdiction of federal and state courts within the Admin State. The visitor to the Website consents to the venue in any action brought against him in connection with breaches of these Terms of Service. The visitor to the Website consents to electronic service of process regarding actions under the above agreement.

For convenience, we have created a page where the sample Terms and Conditions of Use can be found going forward. Although we urge you to seek legal counsel in drafting your terms and conditions, you are welcome to borrow, copy, or link to the sample Terms and Conditions of Use for use on your own site.

http://www.unspam.com/noemailcollection

2.2 Meta Tag for Each Web Page

In addition to your Terms and Conditions of Use page, we are encouraging the widespread adoption the following “Meta Tag” in the header portion of your web pages. To adopt this Meta Tag successfully, website administrators should include this Meta Tag on every page of their respective sites. If you have your own Terms of Service page, replace “http://www.unspam.com/noemailcollection” in the example below with the URL of your own page.

```
<meta name="no-email-collection" content="http://www.unspam.com/noemailcollection" />
```

2.3 Specification for "robots.txt" File

Finally, we suggest you create or modify a robots.txt file for your website. You should include a file called "robots.txt" [SEW] in the root directory of your website. This file dictates how non-human visitors should behave on your website. After any other robots.txt declarations include the following two lines. If you have your own Terms and Conditions of Use page, replace "http://www.unspam.com/noemailcollection" in the example below with the web address of your own Terms and Conditions of Use webpage.

```
User-agent: *
No-Email-Collection: http://www.unspam.com/noemailcollection
```

If email addresses are harvested from your site, then it is arguably a violation of the Terms and Conditions of Use that applies to your page, and you may have a colorable claim for a breach of contract.

2.4 Email Honey Pots for Verification of Compliance

Unfortunately, even with a No-Email-Collection Flag included on your website and appropriate Terms and Conditions of Use in place, it is virtually impossible currently to determine when a spammer has used a spambot to harvest email addresses from your website. To address this problem, we suggest you post on your website a "honey pot" address that you do not use in any email, neither incoming nor outgoing. If these addresses receive email, you have a record that they must have been extracted illegally from your website. The use of several honey pot addresses on a website can help strengthen your case.

A more sophisticated honey pot address could also include a timestamp, the IP address of the visitor to the page, or the session key of the visitor to the page. Including a unique identifier in a honey pot email address allows you to track any users violating your terms of service.

We believe the Terms and Conditions of Use page should contemplate the use of honey pot addresses. The language below gives you a sense as to what such language could look like. If we were using the Terms and Conditions of Use section set forth above, the following text would follow it, but appear before any honey pot addresses. This text will be available at the webpage listed above for the sample Terms and Conditions of Use.

> **RECORDS OF VISITOR USE AND ABUSE**
>
> As a visitor to the Website, you consent to having your Internet Protocol address recorded. The email address immediately below (the "Identifier") is uniquely matched to your Internet Protocol address. Visitors agree not to use this address for any reason.
>
> VISITORS AGREE THAT HARVESTING, GATHERING, STORING, TRANSFERRING TO A THIRD PARTY OR SENDING ANY MESSAGE(S) TO THE IDENTIFIER CONSTITUTES AN ACCEPTANCE AND SUBSEQUENT BREACH OF THESE TERMS AND CONDITIONS OF USE.

2.5 Using Unique Email Honey Pot Addresses

Ideally, it should be possible to trace the usage of an Email Honey Pot address back to the spambot that harvested the address from the website. Such an approach would require a virtually unlimited supply of Email Honey Pot

addresses, one for each spambot or legitimate access to the webpage. The Email Honey Pot address could itself encode the IP address of the visitor to the page, and other information as desired. For example, an IP V4 address is 32 bits long. Using the characters A through Z and zero through 5, each alphanumeric position in the Email Honey Pot address could encode 5 bits of the IP V4 address. So only seven alphanumeric characters would be sufficient to encode an IP V4 address. (A similar approach can be used to encode IP V6 addresses, but it takes 26 alphanumeric characters.) However, the same seven alphanumeric characters should not be used repeatedly for the same IP V4 address, lest the spambots be easily programmed to recognize that address and exclude it. Therefore, some obscuring of the address is required, as long as the process is reversible.

An alternative way to implement the Email Honey Pot address is to randomly but uniquely[1] generate addresses from certain domains (or subdomains). When a random address is generated, an entry is recorded in a database keyed by the Email Honey Pot address that contains the IP address, session key, timestamp, and any other information desired. When an email is sent to an Email Honey Pot address, then the database is consulted for that address to determine how and when it was issued.

2.6 Implementation of an Email Honey Pot

Email Honey Pot addresses can be segregated from legitimate email addresses by using separate subdomain(s) for the Email Honey Pot. For example, the domain holder for EXAMPLE.COM assigns CATCHEM.EXAMPLE.COM to the Email Honey Pot. Each webpage of the EXAMPLE.COM website will contain a link to the Terms and Conditions of Use page for that site, and some or all of the webpages of the website contain a unique Email Honey Pot dynamically generated when that webpage is accessed. An Email Honey Pot address should appear at least on the Terms and Conditions of Use webpage. Appropriate segmentation techniques should be used so the Email Honey Pot address is *not* cached, while the rest of the webpage can be cached if that is otherwise desirable. Also, the use of fonts and colors can make the Email Honey Pot address visible to a spambot while harder to see for a human. For example, the Email Honey Pot address can be used in a "mailto" link from a graphic element located where a human user would be unlikely to click, but a spambot would find it difficult to differentiate.

The domains or subdomains assigned to the Email Honey Pot are assigned separate SMTP mail servers by using the MX record. Thus, spam destined for the Email Honey Pot addresses do increase the load of the SMTP mail servers for legitimate email. Because *all* email destined for Email Honey Pot addresses is spam, the SMTP servers for the Email Honey Pot domains or subdomains can perform actions different than an ordinary SMTP server. The destination Email Honey Pot address can be looked up in the database to find out when it was generated for the website and for whom. That information can be used to sue the spammer, as described below. Furthermore, the contents and path of the spam received at an Email Honey Pot domain can be used to filter spam sent to other domains. Depending on the configuration of the mail system, email similar to spam newly received at an Email Honey Pot domain can be scrubbed from the mailboxes of users that is waiting to be downloaded. Even if the similar email was already downloaded, it could be purged if, for example, the email client uses IMAP.

Multiple subdomains for a domain should be assigned to the Email Honey Pot, lest spammers identify particular domains as spam bait and filter them out. In particular, once a subdomain is used for a lawsuit, it will become known and others will be needed. Another way to deal with this problem is to pool Email Honey Pot subdomains among multiple websites. Even if one Email Honey Pot subdomain becomes known, a website is protected against harvesting because addresses from other Email Honey Pot subdomains are also generated.

Email Honey Pot addresses are easiest to add to dynamically generated websites, such as those using PHP, Perl, ASP, JSP, etc. It is also possible to add Email Honey Pot addresses through an Apache server-side include (SSI),

---

[1] If a duplicate Email Honey Pot address is generated, that fact can be detected by trying to insert the record into the database before returning the contents of the webpage to the requestor. The duplicate insertion will fail, and a new random Email Honey Pot address is generated instead.

however using SSI will require you to run an additional utility script on the server. As a result, you will need the authority to install executable programs on the server hosting your site.

Each website that uses an Email Honey Pot address could maintain its own database and generator of Email Honey Pot addresses. Alternatively, a common database and generator of Email Honey Pot addresses could be shared among multiple websites and used internally by websites that are to include Email Honey Pot addresses. Such a common database and generator could even be a common resource available for any website owner.

3 Legal Basis and Case Law

Courts generally accept that a visitor to a website may be bound by a Terms and Conditions of Use agreement. [PVG] In order to be bound, courts have required the terms to be clearly posted. The precedent as to whether such agreements can apply to non-human visitors, such as spambots, is unresolved. However, by placing a Meta Tag at the beginning of every website as well as in the robots.txt file, arguably non-human visitors are on clear notice. Users of spambots that intentionally ignore these notices arguably should face liability equal to that which they would be exposed to if they personally agreed to the terms. Moreover, the inclusion of an Email Honey Pot address tied to the particular visitor is likely to significantly increase the likelihood your Terms and Conditions of Use will be found to be enforceable against a particular visitor. [PVG] Courts have almost universally held that clicking on a link is sufficient to bind a visitor to an agreement. While unresolved by any court, it seems likely a spambot harvesting your Email Honey Pot address would effectively meet this requirement.

Around the world, governments have recognized individuals' rights to exclude visitors from their property. In an off-line example, sales people are required to respect "no solicitation" signs in many jurisdictions. In the United States, the Supreme Court examined the constitutionality of these off-line laws and has repeatedly upheld them. [FVS] Similar laws are enforced in many additional countries, including: Australia, France, Hong Kong, the Netherlands, and the United Kingdom. [CM] Although there have been some cases on this subject, courts have not settled the issue of exclusion in the electronic world under U.S. Common Law.

Finally, in the United States the CAN-SPAM Act specifically prohibits harvesting of email addresses from websites. The Act makes illegal for anyone to initiate the transmission of an email to a recipient if:

> the electronic mail address of the recipient was obtained using an automated means from an Internet website or proprietary online service operated by another person, and such website or online service included, at the time the address was obtained, a notice stating that the operator of such website or online service will not give, sell, or otherwise transfer addresses maintained by such website or online service to any other party for the purposes of initiating, or enabling others to initiate, electronic mail messages. [CSA]

CAN-SPAM generally does not allow for a private right of action. However, in any breach of contract action by a United States-based website to enforce the No-Email-Collection Flag, plaintiffs can assist enforcement of CAN-SPAM by providing information on violators and encouraging additional legal action against them by ISPs, the Federal Trade Commission and the Department of Justice.

4 Legal Disclaimer

The information in this proposal should not be taken to constitute legal advice. This proposal does not establish any sort of attorney-client relationship between you and the authors of this proposal. It is important to note that this

proposal may not be appropriate in all cases. You should seek your own competent legal counsel when creating terms of service or other legal contracts for your website.

5 Related Work

A similar proposal was suggested for a No Soliciting SMTP Service Extension. [CM] The proposal included broadcasting a no-solicitation flag during the SMTP handshake. Originally proposed by the Coalition Against Unsolicited Commercial Email (CAUCE), and expanded in March 2004 by Carl Malamud, that proposal is complementary with a No-Email-Collection Flag. It may be possible to modify the Terms of Service agreement in order to provide a private right of action, based on breach of contract, for the SMTP proposal as well.

In addition, the Pew Center for Democracy and Technology published a paper [CDT] in March 2003 describing a 6-month research report determining the source of email addresses used for spam. One technique they found effective is the use of encoding email addresses [WBW]. The research found that this encoding could be accomplished in a number of ways as follows.

> Both addresses in the following pairs appear exactly the same to human beings using a modern browser. However, the Pew study found spambots could not read the later address in each pair. The later addresses use ASCII encoding codes to obscure at least the @ sign in the email address.
>
> alan@autos.com → alan@autos.com
> bob@beaches.net → bob@beaches.net
> carl@cat.org → car&#108@cat.org
>
> Mailto anchor tags may also be encoded and still allow a clickable email address to be included in a website without the addresses being exposed to the current generation of spambots. For example, the anchors on the following two lines behave exactly the same when read by a modern browser as part of an HTML document, but the second is resistant to spambots.
>
> email
> email

Unfortunately, while these methods are effective today at stopping spambots, it is inevitable that spambots will adapt as these methods are implemented widely.

References

[BW] Jane Blank, “Unholy Matrimony: Spam and Virus,” Business Week, <http://www.businessweek.com/technology/content/aug2003/tc20030812_7863_tc047.htm>, August 12, 2003, accessed on April 14, 2004.

[CDT] Center for Democracy and Technology, “Why Am I Getting All This Spam? Unsolicited Commercial E-mail Research Six Month Report,” <http://www.cdt.org/speech/spam/030319spamreport.shtml>, March 2003, accessed on April 14, 2004.

[CM] Carl Malamud, “A No Soliciting SMTP Service Extension,” <http://www.ietf.org/internet-drafts/draft-malamud-no-soliciting-07.txt>, March 21, 2004, accessed on June 25, 2004.

[CSA] CAN-SPAM Act, <http://www.spamlaws.com/federal/108s877.html>, Public Law No: 108-187, December 2003, accessed on April 18, 2004.

[FTC] Federal Trade Commission, "Email Address Harvesting: How Spammers Reap What You Sow," <http://www.ftc.gov/bcp/conline/pubs/alerts/spamalrt.htm>, November 2002, accessed on April 18, 2004.

[FVS] See, for example, *Frisby v. Schultz*, 487 U.S. 474 (1988). In *Frisby*, The court upheld the right of a local community to enforce no-solicitation signs posted by individuals and businesses. See also *Rowan v. U.S. Post Office Dep't*, 397 U.S. 728 (1970). In Rowan, the Court allowed individuals to list their postal addresses as off-limits to particular mailers and to allow the post office to assist in enforcing these requests.

[PVG] See, for example, *Pollstar v. Gigmania, Ltd.,* 170 F. Supp. 2d 974, 981 (E.D. Cal. 2000). See also *Specht v. Netscape Communications Corp.,* 306 F.3d 17, 25 (2d Cir. 2002); ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1450 (7th Cir. 1996).

[SEW] Search Engine World, "Robots.txt Tutorial," <http://www.searchengineworld.com/robots/robots_tutorial.htm>, copyright 1996-2002, accessed on April 14, 2004.

[WBW] West Bay Web, "Email Address Encoder," <http://www.wbwip.com/wbw/emailencoder.html>, March 25, 2004, accessed on April 14, 2004.