

**The Secret Lives of Robots.txt**

-

**Sanctioning the Use of Robot  
Exclusion Protocols**

Cyberscholars Group @ Yale-ISP

13 November 2008

Joris van Hoboken, LL.M., M.Sc.

# Affiliations:



**Institute for Information Law**

Doctoral researcher in Law

Dissertation about freedom of expression and search engine law.



**Berkman**

The Berkman Center for Internet & Society  
at Harvard University

**Berkman Center @ Harvard**

**Visiting Researcher Fall 2008**

# Outline

What is robots.txt + short history?

Discussion between search engines and webmasters

Examples of legal sanctioning (in US and EU)

Freedom of expression and information perspectives.

Discussion

## General references:

Niva Elkin-Koren, Let the Crawlers Crawl: On Virtual Gatekeepers and the Right to Exclude Indexing, 26 University of Daytona Law Review 179 (2001);

James Grimmelman, The Structure of Search Engine Law, Iowa Law Review (2008);

Miquel Peguera Poch, When the Cached Link is the Weakest Link: Search Engine Caches under the Digital Millennium Copyright Law, Columbia Law School Public Law & Legal Theory Working Paper 08-176 (2008);

Jonathan Zittrain, The Future of the Internet and How to Stop It, New Haven: Yale University Press, 2008.

## **What is robots.txt?**

Text file, named robots.txt

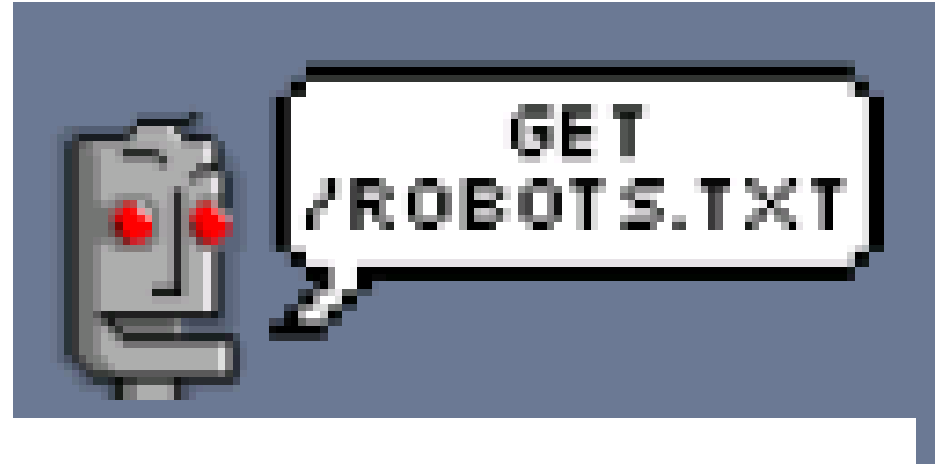
Stored in the root directory of a website

Used to instruct bots/spiders/crawlers which directories of a website not to index.

**The opt-out option major search engines provide to webmasters for indexation.**

# The Web Robots Pages

@ <http://www.robotstxt.org/>



## The Web Robots Pages

Web Robots (also known as Web Wanderers, Crawlers, or Spiders), are programs that traverse the Web automatically. Search engines such as [Google](#) use them to index the web content, spammers use them to scan for email addresses, and they have many other uses.

On this site you can learn more about web robots.

- [About /robots.txt](#) explains what /robots.txt is, and how to use it.
- The [FAQ](#) answers many frequently asked questions, such as "How do I stop robots visiting my site?" and "How can I get the best listing in search engines?"
- The [Other Sites](#) page links to external resources for robot writers and webmasters.
- The [Robots Database](#) has a list of robots.
- The [/robots.txt checker](#) can check your site's /robots.txt file and meta tags.
- The [IP Lookup](#) can help find out more about what robots are visiting you.

# The robots.txt of www.yale.edu

@ <http://www.yale.edu/robots.txt>

```
User-agent: *  
Disallow: /engineering/  
Disallow: /webmaster/stats/  
Disallow: /webmaster/logs/  
Disallow: /napster/  
Disallow: /resnet2008/  
Disallow: /its/software/oracle/  
Disallow: /adminsys/  
Disallow: /search/2003Directory_ORG.pdf  
Disallow: /search/directory_pdfs/DirectoryofOrganizations_05.pdf  
Disallow: /its/communications/attachment-docs/  
Disallow: /grants/findingopps/  
Disallow: /mecha/  
Disallow: /rumpus/  
Disallow: /its/connect/  
Disallow: /yaletomorrow/pdfs/ccmeeting100408.pdf
```

# A discriminatory robots.txt of the Dutch Ministry of Economic Affairs @ <http://www.ez.nl/robots.txt>

```
User-agent: *  
Disallow: /gvscripts  
Disallow: /beheer
```

```
User-agent: Googlebot  
Crawl-delay: 10
```

```
User-agent: Slurp  
Crawl-delay: 10  
Disallow:
```

```
User-Agent: msnbot  
Crawl-delay: 50  
Disallow:
```

```
User-agent: *  
Crawl-delay: 100
```

## **Bias in search engines due to discriminatory robots.txt**

Yes:

-Y. Sun, Z. Zhuang, I. G. Councill, and C. L. Giles. Determining bias to search engines from robots.txt. In Proc. of Int. Conf. on Web Intel. (WI), pages 149–55. IEEE/WIC/ACM, 2007.

-Y. Sun, Z. Zhuang, and C. L. Giles. A large scale study of robots.txt. In Proc. of Int. Conf. on World Wide Web (WWW), pages 1123–4. ACM, 2007.

No (counter research of Yahoo):

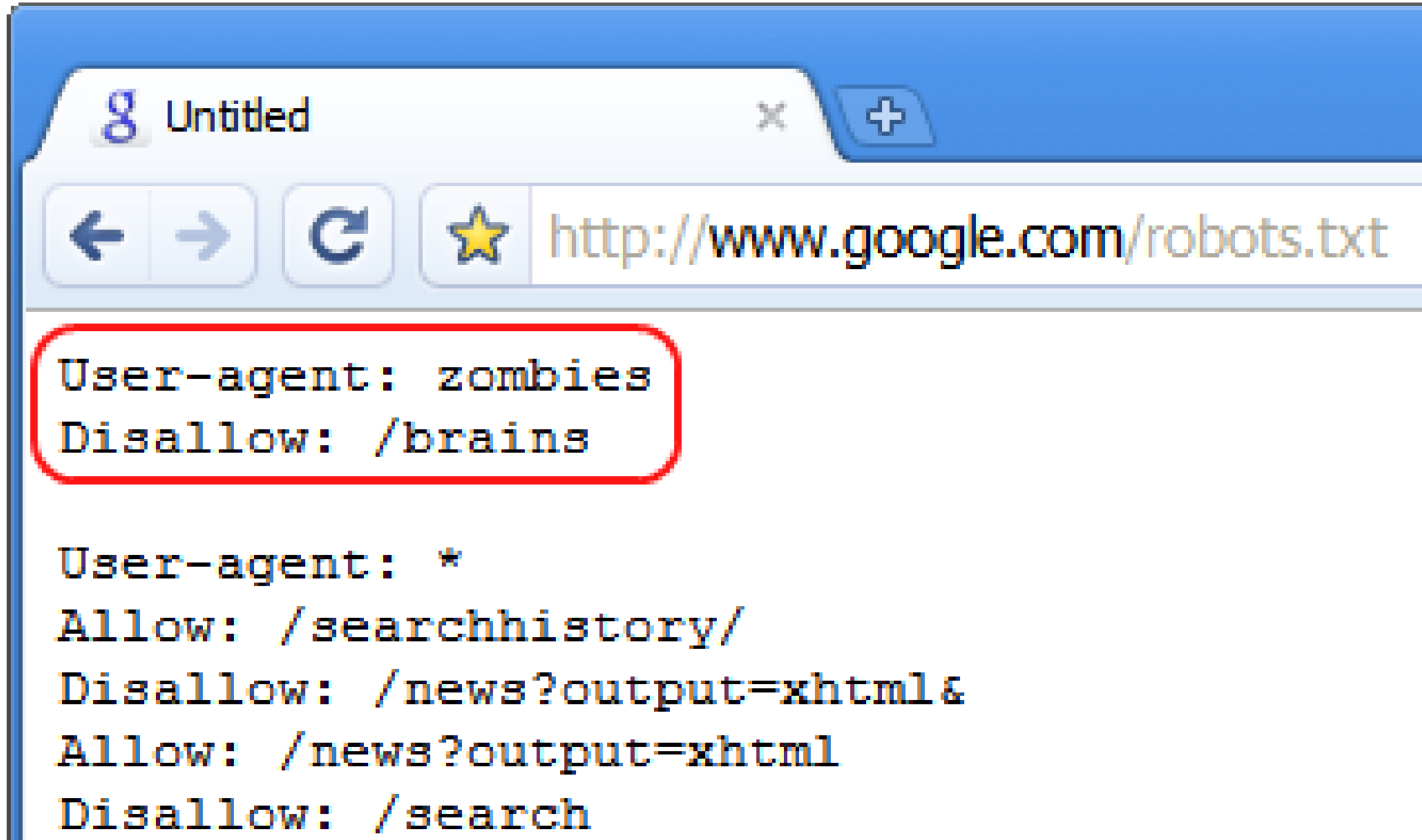
-Santanu Kolay, Paolo D'Alberto, Ali Dasdan, and Arnab Bhattacharjee (Yahoo! Inc.), A Larger Scale Study of Robots.txt, WWW8, 2008.

# The robots.txt file of the White House, excluding 2000+ directories @ <http://www.whitehouse.gov/robots.txt>

```
User-agent:      *
Disallow:       /cgi-bin
Disallow:       /search
Disallow:       /query.html
Disallow:       /omb/search
Disallow:       /omb/query.html
Disallow:       /expectmore/search
Disallow:       /expectmore/query.html
Disallow:       /results/search
Disallow:       /results/query.html
Disallow:       /earmarks/search
Disallow:       /earmarks/query.html
Disallow:       /help
Disallow:       /360pics/text
Disallow:       /911/911day/text
Disallow:       /911/heroes/text
Disallow:       /911/messages/text
Disallow:       /911/patriotism/text
Disallow:       /911/patriotism2/text
Disallow:       /911/progress/text
Disallow:       /911/remembrance/text
Disallow:       /911/response/text
Disallow:       /911/sept112002/text
Disallow:       /911/text
Disallow:       /ConferenceAmericas/text
Disallow:       /GOVERNMENT/text
Disallow:       /QA-test/text
Disallow:       /aci/text
```

# A robots.txt joke from Google

(Source Matt Cutts, Halloween 2008)



```
User-agent: zombies  
Disallow: /brains  
  
User-agent: *  
Allow: /searchhistory/  
Disallow: /news?output=xhtml&  
Allow: /news?output=xhtml  
Disallow: /search
```

## A robots.txt joke about Google



## Short history of robots.txt

The robots.txt was developed in 1993, shortly after the release of the World Wide Web, by Martijn Koster, a Dutch software engineer.

It was discussed on several W3 mailing lists during the mid-nineties.

It is not an official Web or Internet standard.



Martijn Koster @  
<http://www.greenhills.co.uk/mak/mak.html>

**Extending robots.txt** to signal suitability of content for different audiences @ <http://lists.w3.org/Archives/Public/www-talk/1995MayJun/0067.html>

...discussion of the **KidCode proposal** .... labelling scheme ... .. the role technologists play in enabling the content-control of the net...

... Martijn Koster and Ronald Daniel have made very cogent arguments about the need to keep access control, labelling, and subject description apart. ... **Martijn's work with Robot exclusion may provide a workable, ready-to-hand solution. ... A very similar solution could be used for the purpose envisioned by KidCode—**... As a very quick, very dirty method, you could simply have two files: **kids.txt** and **adults.txt**, which advertise either that the site is meant for children or meant for adults; a better solution, I believe, even in the interim, would be an "**audience.txt**", which lists which part of the site contains information appropriate to which audiences. ... Ultimately, of course, it's a hack, and it would need to be replaced.

## A Google webmaster strike proposal @

<http://www.silverspike.co.uk/2007/04/27/bringing-down-google-with-two-simple-lines-of-code/>

### Bringing Down Google With Two Simple Lines of Code

Is Google too powerful? It's a question asked by many. But much of Google's future depends on two simple lines of code.

[...]

One thing they never mentioned (explicitly anyway) was "Site owners continue to give us permission to crawl and index their sites". Without that permission, a large part of Google's business model disappears.

The permission can be taken away with two simple lines of code placed in a site's [robots.txt file](#):

```
User-agent: Googlebot  
Disallow: /
```

## Matt Cutts' poll about NOINDEX-tag

@ <http://www.mattcutts.com/blog/google-noindex-behavior/>

### Some middle ground in between

The vast majority of webmasters who use NOINDEX do so deliberately and use the meta tag correctly (e.g. for parked domains that they don't want to show up in Google). Users are most discouraged when they search for a well-known site and can't find it. What if Google treated NOINDEX differently if the site was well-known? For example, if the site was in the Open Directory, then show a reference to the page even if the site used the NOINDEX meta tag. Otherwise, don't show the site at all. The majority of webmasters could remove their site from Google, but Google would still return higher-profile sites when users searched for them.

### What do you think?

That's the internal discussion that we've been having about NOINDEX meta tags. Now I'm curious what you think. Here's a poll:

### How should Google treat the NOINDEX meta tag?

- Don't show a page at all
- Find some middle ground
- Show a link to the page

## 'legal' comments @ <http://www.mattcutts.com/blog/google-noindex-behavior/>

**EGOL Said,**

February 24, 2008 @ [9:38 pm](#)

I think it could be compared to a "no trespassing" sign on real estate.

**Ian McAnerin Said,**

March 2, 2008 @ [7:02 pm](#)

Regarding the "no-trespassing" sign (which I think is a good analogy), if I put a no trespassing sign on my yard does that mean that the city should remove my address from it's records? Should there be a blocked out space on all maps where my yard is? Or does it mean - stay off my property?

If Google indexes a link to a page on your site from MY site, then that link and it's anchor text is MY content, not yours. Therefore they are allowed to index that content.

**Multi-Worded Adam Said,**

March 3, 2008 @ [8:19 am](#)

Again, everyone's looking at robots.txt vs. the meta robots tag as if it's the real problem. It's not. It only exposes a deeper problem, and a much simpler solution.

The deeper problem: that bot methodology is "opt-out".

The solution: make indexing "opt-in, with opt-out options".

**Dave (original) Said,**

March 4, 2008 @ [9:13 pm](#)

It's a question of ethics and Webmasters rights, IMO.

# **Legal Sanctioning**

# Legal Sanctioning - Property Rights (US)

Ignoring robots.txt exclusion can lead to trespassing claim.

eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000), strong sanctioning.

Intel Corp. v. Hamidi, 71 P.3d 296, 308–09 (Cal. 2003)), weaker property right.

Not resolved, but probably no claim if unauthorized crawling cannot be shown to be egregiously burdensome.

# Legal Sanctioning - Computer Intrusion (US)

If unauthorized computer access in federal and state laws against computer abuse are interpreted broadly, disallow in robots.txt could be seen as making access of bots unauthorized.

For a critical discussion see Orin Kerr, Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes, 78 N.Y.U. L. REV. 1596 (2003).

# Comp. Legal Sanctioning – Database law (EU)

Is not respecting robots.txt circumvention of technological protection?

Dutch Court of Appeal concluded robots.txt is not an effective technological protection measure of a database.

“Even if it is true that other search engines respect measures such as robot[s].txt (which do not prevent but merely request), it does not follow that ZAH [ the search engine in question] by not respecting this code or etiquette, would act unlawfully with regard to the websites.” Court of Appeal Arnhem, 4 July 2006, Zoekallehuizen.nl v. NVM

# Legal Sanctioning – Contract Law

Could robots.txt be interpreted as a part of binding contractual terms for accessing a website?

In the U.S., this might be the case.

# Legal Sanctioning – Copyright (US)

Is indexing and caching infringing?

By using a robots.txt and not telling Google not to index and cache copies through no archive metatags, Field gave Google an implied license to display cached copies. Field v Google - 412 F.Supp.2d 1106 (2006). Case does not resolve initial copies.

Parker v. Yahoo, 07-0257 District Court Penns. 2008: Also concludes that not using robots.txt and no archive to exclude material from SEs and their cached copies archive, give SEs implied license.

# Legal Sanctioning – Copyright (EU)

Does the opt-out of robots.txt resolve copyright problems between rights holders and search engines?

Not yet resolved at **EU** level.

National peculiarities apply.

# Legal Sanctioning – Copyright (Germany)

Some German courts have concluded that not using robots.txt to exclude material implied a license.

Recently, in the context of image search, LG Hamburg , 26.09.2008  
- Az.: 308 O 42/06:

"1. Durch das Online-Stellen von Bildern auf seiner Webseite erteilt der Webseiten-Betreiber Google Inc. keine Einwilligung, urheberrechtlich geschützte Bilder als automatische Thumbnails anzuzeigen.

**2. Die Einwilligung ergibt sich auch nicht daraus, dass es ein Webseiten-Betreiber durch entsprechende Maßnahmen ("robots.txt". ".htaccess") in der Hand hat, die Öffentlichkeit oder Teile der Öffentlichkeit von der Nutzung seiner Webseite auszuschließen. Internationale Standards z.B. des World Wide Web Konsortiums W3C oder des Robots Exclusion Standard Protocols sind für die rechtliche Beurteilung unverbindlich."**

# Legal Sanctioning – Copyright (Belgium)

Copiepresse v. Google dispute about Google News

Copyright online is opt-in, not opt-out, also for search engines:

Attendu que comme souligné dans les conclusions déposées par la Sofam, le droit d'auteur n'est pas un droit d'opposition mais un droit d'autorisation préalable ; Que cela signifie que l'autorisation doit être obtenue de manière certaine, préalablement à l'utilisation envisagée ;

Copiepresse v. Google, TGI Bruxelles, 13 Feb 2007;  
[http://www.copiepresse.be/copiepresse\\_google.pdf](http://www.copiepresse.be/copiepresse_google.pdf)

# From a footnote in the Green Paper on the Knowledge Economy (EU)

*“[S]earch engines are not asking for prior permission from copyright owners to index content of web pages. Search engines argue that, if a content owner does not want the content of the web page to be indexed, he can encode the message in a text file called "robots.txt" in order to opt-out and block the search engine from copying content. If no such technology is applied, they believe that this is tantamount to an implied licence for a search engine to copy and index.”*

**Implicitly: search engines might need permission of websites to crawl and index content.**

# Automated Content Access Protocol (ACAP)

Recent global publishers initiative

Builds on the idea that robot instruction protocols can be used for licensing.

It is an extension of robots.txt, making more granular instructions of publishers to search engines possible.

APAC gives publishers the option to

- allow and disallow crawling for parts of their content;
- and for certain usages after crawling;
- and setting a time frame for indexing and indexed content.

# ACAP Usage Types

present-original

present-currentcopy

present-oldcopy

present-snippet

# ACAP robots.txt of major Dutch newspaper @ [www.ad.nl/robots.txt](http://www.ad.nl/robots.txt)

```
##ACAP version=1.0

# Robots.txt voor AD

# REP

# Algemene instellingen
User-agent: *
Disallow: /training/

# Uitzonderingen
User-agent: Mozilla/4.0 (compatible; HowardsHome/1.0; +http://www.howardshome.com)
Disallow: /
User-agent: Nutch
Disallow: /

# ACAP

# Algemene instellingen
ACAP-crawler: *
ACAP-disallow-crawl: /training/
ACAP-disallow-preserve: /
ACAP-disallow-present-snippet: /
ACAP-allow-index: *.html$ time-limit=14-days
ACAP-allow-index: $ time-limit=1-days
ACAP-allow-index: /$ time-limit=1-days

# Uitzonderingen
ACAP-crawler: Mozilla/4.0 (compatible; HowardsHome/1.0; +http://www.howardshome.com)
ACAP-disallow-crawl: /
ACAP-crawler: Nutch
ACAP-disallow-crawl: /
```

## Legal Sanctioning – Safe Harbour CDA 230 (US)

Orin Kerr @ Volokh Conspiracy: *“I'm no expert ... [b]ut here's my amateurish idea: **Would it help to somehow link up provider immunity with search robot exclusion?** Under current law, site owners are immune from liability for the speech of others under 47 U.S.C. 230. This means that a site owner can allow anonymous comments, announce that anything goes, and then sit back and watch as the trolls engage in all sorts of foul play. Search engine robots then pick up the foul play, resulting in harm weeks or months later when a third party googles that person or event. A lot of people may be harmed, but the law can't stop it: the provider is immune and the commenters are anonymous.”*

# Legal Sanctioning – Data Privacy (EU)

Art. 29 Working Party: *“Website owners may opt out a priori of both the search engine and the caching by using the robots.txt file or the Noindex/NoArchive tags. This may be more than an optional solution. Publishers of personal data need to consider whether their legal basis for publication includes indexing of this information by search engines, and create respective safeguards as necessary, including, but not limited to, use of the robots.txt file and/or Noindex/NoArchive tags. It is essential that search engine providers respect opt-outs expressed by website editors. This opt-out can be expressed before the first crawling of the website or once it has already been crawled; in that case, updates on the search engine should be carried out as soon as possible.”*

See Article 29 Working Party, Opinion 1/2008 on data protection issues related to search engines, April 2008, [http://ec.europa.eu/justice\\_home/fsj/privacy/docs/wpdocs/2008/wp148\\_en.pdf](http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2008/wp148_en.pdf)

See also: Dutch Data Protection Authority, Guidelines Publication of Personal Data on the Internet, Dec 2007, [http://www.dutchdpa.nl/downloads\\_overig/en\\_20071108\\_richtsnoeren\\_internet.pdf?refer=true&theme=purple](http://www.dutchdpa.nl/downloads_overig/en_20071108_richtsnoeren_internet.pdf?refer=true&theme=purple)

# **Freedom of Expression and Information**

# Freedom of Expression and Information

Freedom to access publicly available/ accessible information

Dominance, pluralism, diversity

Findability (analog analogies of control over findability?)

Effective access to publicly available/accessible information in the era of information overload.

# Questions and Comments

Joris van Hoboken, LL.M., M.Sc.

Institute for Information Law

E-mail: [vanhoboken@ivir.nl](mailto:vanhoboken@ivir.nl)

Blog: [www.jorisvanhoboken.nl](http://www.jorisvanhoboken.nl)

# For discussion

1. Legal sanctioning of robots.txt should not reinforce dominance of one particular search engine.
2. There are too many different general laws applicable to crawling by search engines. Stifles innovation.
3. Robots.txt can be used illegitimately.
4. Robot exclusion and crawling should be facilitated through regulation.
5. The law should give online publishers and website owners complete/some/no control over findability of publicly accessible material on the Internet.
6. There should be common on public interest crawlers and index with special legal privileges.

# Robots.txt can also be a blog

@ <http://www.webmasterworld.com/robots.txt>

```
# Here a Bot - There a Bot - Every Where a Bot Bot #
# by brett tabke #
# 12/19/2005 #
# #
# ===== #

# After alot of testing and bot busting, the current robots.txt is what was
# settled on. I felt exposing the code was the best way to explain it all
# (see the actual robots.txt above for the full story).
#
# Testing the bots code and the security code to get it all right took alot
# of time. In the end, we found:
#
# - A surprising 21 bots that were following all the active list posts on a
#   daily basis and downloading that content.
#
# - About 45 trademark and other page monitoring services. The majority of
#   those monitoring services obey robots.txt.
#
# - 15 bots would accept cookies.
#
# - 2 more web sites reselling WebmasterWorld content. One in China and one
#   in the stans. both out of legal reach.
#
```

In 2007, on SES, Yahoo proposed changes that would allow webmasters to signal no crawling of parts of a web page  
@ <http://www.ysearchblog.com/archives/000444.html>

MAY 02, 2007

## Introducing Robots-Nocontent for Page Sections

We recently returned from our annual rendezvous at [SES New York](#) and, like always, learned a lot from our webmasters. The ['Robots.txt Summit'](#) generated some healthy discussions and support for adding a tag to parts of a page that do not relate to the main content, such as navigation, menus repeated across the entire site, boilerplate text, or even advertising. We heard what people were asking for so we did a little homework and are now happy to introduce the 'robots-nocontent' tag.

This tag is really about our crawler focusing on the main content of your page and targeting the right pages on your site for specific search queries. Since a particular source is limited to the number of times it appears in the top ten, it's important that the proper matching and targeting occur in order to increase both the traffic as well as the conversion on your site. It also improves the abstracts for your pages in results by omitting unrelated text from search result summaries.

To do this, webmasters can now mark parts of a page with a ['robots-nocontent' tag](#) which will indicate to our crawler what parts of a page are unrelated to the main content and are only useful for visitors. We won't use the terms contained in these special tagged sections as information for finding the page or for the abstract in the search results. Note: Using a "nocontent" tag to mark explicit sections of content is not considered "cloaking" because all of the content on the page is available to protect the relevance of the results (unlike "cloaking" where we may be served content that is different from what visitors see).