Core - Feature # 14798

Status:	Accepted	**Priority:**	Should have
Author:	Jody Cleveland	**Category:**	Indexed Search
Created:	2005-06-06	**Assigned To:**	
Updated:	2015-06-15	**Due date:**	
PHP Version:	5.5		
Complexity:			
Sprint Focus:			
Subject:	Robots.txt and indexed search		

Description

I've got a handful of those pdf files that I don't want
indexed. So, to satisfy google, and other search engines, I use a
robots.txt file so they aren't indexed.

My requested feature is, getting typo to honor robots.txt files, and
skip indexing files listed in there.

(issue imported from #M1170)

History

#1 - 2005-06-14 08:32 - Michael Stucki

Are you talking of internal or external files?

Let's say your site is http://www.mysite.com/mysitedir/ and you have several links, which of them do you think should be checked against Robots.txt?

http://www.mysite.com/mysitedir/fileadmin/test.pdf
http://www.mysite.com/mysitedir/Intro.5.0.html
http://www.mysite.com/anothersitedir/fileadmin/test.pdf
http://www.mysite.com/anothersitedir/Introl.5.0.html
http://www.anothersite.com/mysitedir/fileadmin/test.pdf
http://www.anothersite.com/mysitedir/Intro.5.0.html

#2 - 2005-06-14 18:27 - Jody Cleveland

I would think anything listed in robots, as long as it was within the site. More like this one:

http://www.mysite.com/mysitedir/fileadmin/test.pdf [^]

#3 - 2005-07-20 01:10 - Michael Stucki

I will implement this if you can do some research for me about Robots.txt:

- Is there an RFC?
- Where is the file expected to be found: Only in / or in any directory of the rootline?
- Does it accept regular expressions or only plain strings?
- Any other special formattings?

#4 - 2005-09-27 16:21 - Jody Cleveland

Here's the RFC:

3.3 Formal Syntax

This is a BNF-like description, using the conventions of RFC 822 [5], except that "|" is used to designate alternatives. Briefly, literals are quoted with "", parentheses "(" and ")" are used to group elements, optional elements are enclosed in [brackets], and elements may be preceded with <n>* to designate n or more repetitions of the following element; n defaults to 0.

robotstxt = *blankcomment
| **blankcomment record *(1*commentblank 1*record)**
***blankcomment**

blankcomment = 1(blank | commentline)
commentblank = *commentline blank *(blankcomment)
blank = *space CRLF
CRLF = CR LF
record = *commentline agentline *(commentline | agentline)
1*ruleline *(commentline | ruleline)
agentline = "User-agent:" *space agent [comment] CRLF
ruleline = (disallowline | allowline | extension)
disallowline = "Disallow" ":" *space path [comment] CRLF
allowline = "Allow" ":" *space rpath [comment] CRLF
extension = token : *space value [comment] CRLF
value = <any CHAR except CR or LF or "#">
commentline = comment CRLF
comment = **blank "#" anychar**
space = 1(SP | HT)
rpath = "/" path
agent = token
anychar = <any CHAR except CR or LF>
CHAR = <any US-ASCII character (octets 0 - 127)>
CTL = <any US-ASCII control character
(octets 0 - 31) and DEL (127)>
CR = <US-ASCII CR, carriage return (13)>
LF = <US-ASCII LF, linefeed (10)>
SP = <US-ASCII SP, space (32)>
HT = <US-ASCII HT, horizontal-tab (9)>

The syntax for "token" is taken from RFC 1945 [2], reproduced here for convenience:

token = 1*<any CHAR except CTLs or tspecials>
tspecials = "(" | ")" | "<" | ">" | "@"
| "," | ";" | ":" | "\" | <">
| "/" | "[" | "]" | "?" | "="
| "{" | "}" | SP | HT

The syntax for "path" is defined in RFC 1808 [6], reproduced here for convenience:

path = fsegment *("/" segment)
fsegment = 1*pchar
segment = *pchar
pchar = uchar | ":" | "@" | "&" | "="

```
    uchar       = unreserved | escape
    unreserved  = alpha | digit | safe | extra
escape      = "%" hex hex
    hex         = digit | "A" | "B" | "C" | "D" | "E" | "F" |
                      "a" | "b" | "c" | "d" | "e" | "f"
alpha       = lowalpha | hialpha
lowalpha    = "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" |
                "j" | "k" | "l" | "m" | "n" | "o" | "p" | "q" | "r" |
                "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z"
    hialpha     = "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" |
                "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" |
                "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z"
digit       = "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" |
                "8" | "9"
safe        = "$" | "-" | "_" | "." | "+"
    extra       = "!" | "*" | "'" | "(" | ")" | ","
```

I belive the robots.txt file needs to be in the root of the site:

This section contains an example of how a /robots.txt may be used.

A fictional site may have the following URLs:

http://www.fict.org/

http://www.fict.org/index.html

http://www.fict.org/robots.txt

http://www.fict.org/server.html

http://www.fict.org/services/fast.html

http://www.fict.org/services/slow.html

http://www.fict.org/orgo.gif

http://www.fict.org/org/about.html

http://www.fict.org/org/plans.html

http://www.fict.org/%7Ejim/jim.html

http://www.fict.org/%7Emak/mak.html

The site may in the /robots.txt have specific rules for robots that send a HTTP User-agent "UnhipBot/0.1", "WebCrawler/3.0", and "Excite/1.0", and a set of default rules:

1. /robots.txt for http://www.fict.org/
2. comments to webmaster@fict.org

```
User-agent: unhipbot
    Disallow: /
User-agent: webcrawler
    User-agent: excite
    Disallow:
User-agent: *
    Disallow: /org/plans.html
    Allow: /org/
    Allow: /serv
    Allow: /~mak
   Disallow: /
```

The following matrix shows which robots are allowed to access URLs:

unhipbot	webcrawler & excite	other
http://www.fict.org/	No	Yes	No

http://www.fict.org/index.html	No	Yes	No
http://www.fict.org/robots.txt	Yes	Yes	Yes
http://www.fict.org/server.html	No	Yes	Yes
http://www.fict.org/services/fast.html	No	Yes	Yes
http://www.fict.org/services/slow.html	No	Yes	Yes
http://www.fict.org/orgo.gif	No	Yes	No
http://www.fict.org/org/about.html	No	Yes	Yes
http://www.fict.org/org/plans.html	No	Yes	No
http://www.fict.org/%7Ejim/jim.html	No	Yes	No
http://www.fict.org/%7Emak/mak.html	No	Yes	Yes

I took all this from this document:
http://www.robotstxt.org/wc/norobots-rfc.html

I hope that helps, and I really appreciate you looking into this. If there's anything else you need, let me know.

#5 - 2014-12-04 16:50 - Mathias Schreiber

- Description updated

- Status changed from New to Accepted

- Target version changed from 0 to 7.0

- PHP Version set to 5.5

#6 - 2014-12-23 19:48 - Mathias Schreiber

- Target version changed from 7.0 to 7.1 (Cleanup)

#7 - 2015-06-15 18:11 - Benjamin Mack

- Target version changed from 7.1 (Cleanup) to 7.4 (Backend)