

Chapter Twelve (g): Hostnames & Regular Expressions (Regex) for Smartfilter and Netsweeper

Summary of Chapter:

- General Information on how Regex is used in Site Lists on the Pilot with Smartfilter and Netsweeper.

What you need:

- Knowledge of Admin user account and valid password for your Pilot.
- A UTM PoP code. NetPilot users can purchase this from the following address: <http://www.equinet.com/ordering/default.asp>, CachePilot users please contact Equinet for a quote.

Software Revision Required:

- Applicable to software revision 5.2.0 > Net/CachePilots
(Net/CachePilot will be referred to as 'Pilot'. All image examples are of a NetPilot.)



Site Lists are part of 'Web access rules' which are applied by editing 'User Groups' in the User Accounts / Groups section. For more information on Sites Lists please see the other sections of Chapter 12.



The below information will not work if you are using Guardian. Please see Chapter 12 (i)



Site Lists & Regex:

- Log on to the Pilot as shown in Chapter One (b).
- From the left-hand side of the screen, select 'Web', then 'Filtering' and then 'Site lists'. (All links are highlighted below).



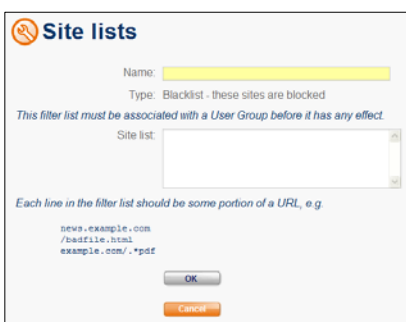
- The default Site Lists are highlighted left.

- Select a Site List you wish to edit, and then select 'OK'.



These Site Lists are blank as default, until you edit them.

- You will be presented with a similar screen to the below:





In the 'Site List:' text box you enter URLs you wish to allow or block. Some of these will require Regex.



In the 'Site List:' text box there is a character size limit of 10,000 characters. Be careful not to exceed this.


Blocking/Allowing particular strings in the Domain Name:

 In the first part of the URL (before the '/') the NetPilot uses pattern match not Regex.
 Therefore the only special characters that you need to use in the first part of the URL are listed below:

- ? - One '?' matches exactly one letter. To match five letters you would need to enter '?????'.
- * - Matches any number of characters




Domain names (www.google.co.uk) are split up into words by '.'s and the domain name (before the '/') part of the pattern only matches whole words not half a word; unless you use a special character like '*' which matches any number of any character. When the Pilot is matching the whole word, it matches from the right to the beginning of the hostname.

 Listed below are examples that you can enter into the Site List text box provided and what URLs the Pilot will match them to.


Examples entered into NetPilot	Examples of URLs that will match	Examples that will <u>NOT</u> match
foo.example.com	www.foo.example.com top.foo.example.com foo.example.com	food.example.com
example.com	wibble.example.com example.com	ample.com
com	example.com www.example.com	co.uk de
???.example.com	foo.example.com biz.example.com	b.example.com ba.example.com
foo.example.*	foo.example.co.uk www.foo.example.com	foo.example

Blocking/Allowing 'hostname:port' URLs:

 The same pattern match rules apply to the below as they are displayed before the first '/'.

- foo.example.com:8000 - matches access to a particular port number
- :8???
- .* - matches any port number on any host

Blocking pure IP Addresses but allowing Domains with Numbers:

 To block a pure IP Address like '144.12.1.97' but still allow domains with numbers e.g. www.example10.co.uk or www.1610985.com you will need to enter in the following:

- *0
- *1
- *2
- *3 - Before the '/' it is matched right to left of the domain, the
- *4 pattern matches the last bit of the domain. So this pattern will
- *5 never match a domain with a number in it as domains always
- *6 end in 'com', 'co.uk' etc.
- *7
- *8
- *9

Blocking/Allowing particular strings after the first '/':



The second part of the URL (the section after the first '/') is case-insensitive 'extended regular expression' (Regex). Regex is a well-documented and is widely used for many things, like program parsers and text processors. For further information please see the Foot notes at the end of this document.

ⓘ Though Regex is widely documented, below is a listed of syntax that you might find useful (below **a** and **b** are representing single characters, **m** and **n** are representing numbers, **s** is representing a set of things and **x** and **y** are representing any character as well as Regex expressions):

.	-	is a wildcard which matches any one character
x*	-	matches any number of optional things matching X
\x	-	turns off the special meaning of X
^	-	matches the notional beginning marker
\$	-	matches the notional end marker
(x y)	-	matches either x or y
x?	-	the letter before the '?' is optional
x+	-	X is required and may be repeated any number of times
[s]	-	match any single character from the set s
[a-b]	-	match any single character from the range 'a-b'
[^s]	-	any single character but not from set s
[^a-b]	-	any single character but not the characters from the range 'n-x'
x{n}	-	x can be repeated up to n times
x{m,n}	-	x can be repeated between m and n times

ⓘ An example is shown below with some of the syntax above:

`.*\mp3(\?|$)` - this will block MP3 music downloads

ⓘ Below explains each part of this example:

/	-	indicates that this is a path-pattern, not a host-pattern
.*	-	matches any number of characters
\.	-	matches a '.' and nothing else, cancels out the '.' special meaning of any character
mp3	-	the text to match in the URL
(\? \$)	-	the pattern can match when the URL ends or ends with a '?'

Blocking/Allowing certain file types in the URL e.g. MP3:

ⓘ To block certain file types in the Regex side of the URL (after the first '/') you need to enter the following:

<code>.*\exe(\? \$)</code>	-	this will block Executable downloads using HTTP
<code>.*\eml(\? \$)</code>	-	this will block the file type used by the W32/Nimda I-Worm

ⓘ If you wanted to block a number of file types you can shorten the strings of syntax to one, saving on your 10,000 character size limit.

`.*\.(exe|eml|mp3)(\?|$)`

ⓘ To block a file type for a certain domain, you will need to enter the following:


`boo.com.*\.(exe|eml|mp3)(\?|$)`
`boo.com.*\exe(\?|$)`

Blocking/Allowing a particular string:




This only applies to the domain name before the first '/' only.



 To find the sequence 'games' anywhere in the hostname would be:

`*games*`

 If you want to apply this rule only to certain country domains it would be :

`*games*.com`


or

`*games*.de`



The next section only applies to the URL after the first '/' only.



 To allow or block a particular string in the right-hand part of the URL, you would enter the following:

`/*games`

Foot notes:

Description of Regex: www.regular-expressions.info/posix.html