

Chapter Twelve (h): Regular Expressions (Regex) for Guardian

Summary of Chapter:

📄 General Information on how Regex is used in Site Lists on the Pilot with Guardian.

What you need:

📄 Knowledge of Admin user account and valid password for your Pilot.

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 Smartfilter or Netsweeper.



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).



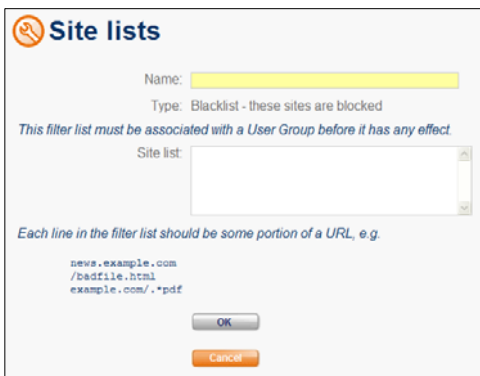
📄 The default Site Lists are highlighted left.

📄 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.



Entering URL's into Site Lists for Guardian into are 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 |




Some useful examples are given below on how to use the above syntax.

Blocking/Allowing particular Domain Names:


ⓘ Listed below are examples that you can enter into the Site List text box provided and what URLs the Pilot can and will not match them to.

| Examples entered into NetPilot | Examples of URLs that will match | Examples that will <u>NOT</u> match |
|--------------------------------|---|--|
| ^www\.bbc\.co\.uk\$ | www.bbc.co.uk | bbc.com bbc.co.uk/news/general |
| bbc\.co\.uk | www.bbc.co.uk www.news.com/info/bbc.co.uk/home | www.bbc.com |
| com | example.com www.example.co.uk/help/doc/comma/ | co.uk de |
| com/ | example.com www.home.com | www.example.co.uk/help/doc/comma/ |
| ^[^/]*bbc\.co\.uk/ | www.bbc.co.uk bbc.co.uk/news/info | www.example.com/info/bbc.co.uk/home bbc.com |


Blocking/Allowing a particular string:

 To find the sequence 'games' anywhere in the URL you would need to enter:

games


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

`^[^/]*games.*\.com` or `^[^/]*games.*\.de`

 To allow or block the word 'games' after the domain name, you would enter the following:


`/. *games`

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:

`^[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}`


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

 To block or allow certain file types within the URL's you need to select the appropriate tick boxes which are given when you edit any Blacklists or Whitelists. A similar screen is shown below:



 If you wish to block an individual file type in the URL you can enter the following:

`.*\exe(?:$)` - this will block Executable downloads using HTTP
`.*\eml(?:$)` - this will block the file type used by the W32/Nimda I-Worm
`.*\mp3(?:$)` - this will block MP3 music downloads

 Below explains each part of the examples above:

`/` - 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 '?'

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

`boo\.com.*\exe(?:$)`

Foot notes:

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