Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 1

I will assume you've looked at Primer #01 first. Otherwise many of the steps done here might not make a lot of sense.

I maintain the majority my MailWasherPro filters in an Excel spreadsheet. The older filters were entered and maintained in MailwasherPro manually. At some point, I will probably re-visit the older filters to determine if any optimizations are possible.

Since I've started developing and using the newer Regular Expression filters that check the header, MailWasher Pro is detecting and deleting between 80% and 90% of the bad guys that get by the spam filter checking on the Mail Servers. My automated and semi-automated Regular Expressions have made life much easier. Designing and documenting what I've done has taken some time, but I'm delighted to not have to see most bad guys (spam) after MailWasher Pro detects and deletes them.

I'm happy to share what I've learned with the MailWasherPro community and I'm happy to receive feedback on what I'm sharing.

1. Another more complex use of RegularExpressions is to deal with a huge number of "fake" domain suffixes that are being used to send out messages:

- Examples:
  - aaaa@bbbb.stream
  - aaaa@bbbb.win
  - aaaa@bbbb.bid
  - aaaa@bbbb.pro (this is a legitimate suffix, but it's usually spam)
  - aaaa@bbbb.cricket

The Filter looks like this:

Header	Contains	RegEx	[.](men|stream|party|date|faith|bid|cricket|xyz|win|moscow|pro)[>)/]

- The brackets at the beginning [.] indicate to look for a period at the start of the sub-string before the "things"
- The () parentheses indicate the groups of "things" to look for
- The | (pipe character) separates different "things" to look for
- The brackets at the end of the substring [>)/] indicates to look for one of these three characters at the end of the sub-string - reason: in a header, an email address appears several times and is followed immediately by one of these three characters where it occurs. We want to find only instances where the "things" in the header are preceded by a period and followed by one of these three characters.

- You can check for valid domain suffixes at: http://www.computerhope.com/jargon/num/domains.htm
  - I note above that .pro is valid, but it is not widely in use yet

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 2

2. Another more complex use of RegularExpressions is to deal with a large number of emails from Dr. Oz where the header data specifies various combinations for Dr. Oz
   - **The key here and henceforth (in the use of regular expressions in this document) is the use of the sub-expression [-_.]*. This is a substitution (whether manually done or automated by an Excel formula) to replace a blank space with the following sub-expression: "[-_.]*". This sub-expression forces the search for alternate separators (dash "-" , underscore "_" , period "." or space " "] or for no separator forced by a "*".**

- Examples:
  - dr oz
  - doctor oz
  - d0ct0r oz
  - etc.

- **The Filter looks like this:**

Header	Contains	RegEx	(dr[-_.]*(o|0)z|d(o|0)ct(o|0)r[-_.]*(o|0)z)[ ]*

- The first set of () parentheses indicate sub groups of "things" to look for
  - The first set looks for all the variants of "dr"
- The first set of brackets [-_.] indicates to look in the first sub group for separators such as:
  - dr-
  - dr_
  - dr.
- The asterisk following the right bracket [-_.]* provides for the case where the dr oz string has no separator
  - dr
- The first inner set of () parentheses indicate sub groups of "things" to look for
  - The pipe | looks for all the variants of "o" – either:
    - o
    - 0
  - The above combination with the letter "z" following the right parentheses indicates the search for :
    - oz
    - 0z
- So the entire expression within the first parentheses indicates a search for all variants of dr oz including:
  - dr-oz
  - dr_oz
  - dr.oz
  - dr oz

  - droz
  - dr-0z
  - dr_0z
  - dr.0z
  - dr 0z
  - dr0z
- The next pipe character | followed by a second set of parentheses indicates a similar search for variants of "doctor oz" including the search for **all variants** such as:
  - doctor-oz
  - doctor_oz
  - doctor.oz
  - doctor oz
  - doctoroz
  - d0ctor-oz
  - doct0r_0z
  - d0ct0r oz
  - d0ct0r0z
  - doct0r.0z
  - and **all other possible variants** of the 0 for o substitution, with and without separators

- At the end of the second sub-string , the brackets with the space and followed by an asterisk [ ]* indicate to look for zero or more spaces following the string.

- **NOTE:** This rule could have been optimized further since the checking for dr or doctor is always followed by the checking for oz. So the simple check could group the "dr" and "doctor" checks together and have only a single check for "oz". This optimized version of the rule also accounts for the substitution of a "2" for a "z".

The Filter looks like this:

Header	contains	RegEx	(dr|d[o0]ct[o0]r)[-_.]*[o0][z2][ ]*

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 4

3. When you have a series of compound **2-word phrases** that are commonly found in tandem and you find multiple instances and combinations of these phrases being used, a rule to catch these phrases would be look something like this:
    - The first word of the phrase is commonly:
        - Asian
        - Russian
        - Baltic
        - Filipina
        - Korean
        - Etc.
    - The second word of the phrase is commonly:
        - Women
        - Escort
        - Bride
        - Honey
        - Lady
        - Etc.
    - So some combinations of the above would be
        - Asian women
        - Asian escort
        - Asian bride
        - Asian honey
        - Asian lady
        - Asian-women
        - Asian_women
        - Asian.women
        - Asianwomen
        - Filipina bride
        - Korean_honey
        - Baltic.spouse
        - European companion
        - Etc.
    - These phrases can be collected in a filter and updated as needed

The Filter looks like this:

Header	Contains	RegEx	((asian|russian|baltic|korean|latin|chinese|ukrainian|asia|filipina|filipino|cambodian|vietnamese|ethnic|european)[-_.]*(Ladi|lady|girl|women|escort|chick|broad|beaut|mate|honey|woman|bride|whore|soulmate|elegant|young|single|wife|spouse|companion))

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 5

4. Where you can really optimize your use of Regular Expressions as filters is to use them in tandem with an MS-Excel spreadsheet, where you can write formulas to create the regular expression strings. Once you've perfected this, you don't have to remember how it's done as Excel does all the work for you. You just edit the formulas and copy/paste the results.

- When you have a series of compound **multi-word phrases** that are commonly found in tandem and you find multiple instances and combinations of these phrases being used. This Example uses 5 instances of these phrases. The final generated Regular Expression (in Column G) will check all possible combinations for a total of 125 instances, with separators [-_.] and without. The phrases do not need to have equal numbers of entries on each side:
  - Type/paste the first part of the phrase into Column A:
    - amazon
    - burger king
    - macys
    - macy's
    - outback steakhouse
  - Type/paste the second part of the phrase into Column D:
    - reward
    - account
    - gift card
    - voucher
    - party planner
  - Let Excel do the work of creating the regular expressions
    - Column A contains the text of the first phrase
    - Column B contains a pipe separator character if Column A is non-blank
    - Column C contains a Regular Expression filter generated by an Excel formula using Columns B and A
    - Column D contains the text of the second phrase
    - Column E contains a pipe separator character if Column D is non-blank
    - Column F contains a Regular Expression filter generated by an Excel formula using Columns E and D
    - Column G contains a concatenated Regular Expression filter from the separate filters in Columns C & F.
  - You setup the formulas once and allow for the number of phrases that you determine – I started out with 35 and then found that I had to go to 70 for each phrase.
  - This example provides for 5 rows of data for each part of the compound phrase. Take care to use the correct cell addresses in Columns C and F and when expanding the formulas.
  - **The Column G value is the regular expression to be pasted into a MWP filter**

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 6

Column A	Row 1	Amazon
	Row 2	Burger king
	Row 3	Macys
	Row 4	Macy's
	Row 5	Outback steakhouse
	Etc.	

Column B	Row 1	=IF(ROW(A1)=1,"|",IF(A2=A1,"*","|"))
	Row 2	=IF(ROW(A2)=1,"|",IF(A3=A2,"*","|"))
	Row 3	=IF(ROW(A3)=1,"|",IF(A4=A3,"*","|"))
	Row 4	=IF(ROW(A4)=1,"|",IF(A5=A4,"*","|"))
	Row 5	=IF(ROW(A5)=1,"|",IF(A6=A5,"*","|"))
	Etc.	

Column C	Row 1 Formula	=IF(B1="|",B1&SUBSTITUTE(A1," ","[-_.]*"),"")&IF(B2="|",B2&SUBSTITUTE(A2," ","[-_.]*"),"")&IF(B3="|",B3&SUBSTITUTE(A3," ","[-_.]*"),"")&IF(B4="|",B4&SUBSTITUTE(A4," ","[-_.]*"),"")&IF(B5="|",B5&SUBSTITUTE(A5," ","[-_.]*"),"")
	Row 1 Value	|Amazon|Burger[-_.]*king|Macys|Macy's|Outback[-_.]*steakhouse

Column D	Row 1	Reward
	Row 2	Account
	Row 3	Gift card
	Row 4	Voucher
	Row 5	Party planner
	Etc.	

Column E	Row 1	=IF(ROW(D1)=1,"|",IF(D2=D1,"*","|"))
	Row 2	=IF(ROW(D2)=1,"|",IF(D3=D2,"*","|"))
	Row 3	=IF(ROW(D3)=1,"|",IF(D4=D3,"*","|"))
	Row 4	=IF(ROW(D4)=1,"|",IF(D5=D4,"*","|"))
	Row 5	=IF(ROW(D5)=1,"|",IF(D6=D5,"*","|"))
	Etc.	

Column F	Row 1 Formula	=IF(B1="|",B1&SUBSTITUTE(A1," ","[-_.]*"),"")&IF(B2="|",B2&SUBSTITUTE(A2," ","[-_.]*"),"")&IF(B3="|",B3&SUBSTITUTE(A3," ","[-_.]*"),"")&IF(B4="|",B4&SUBSTITUTE(A4," ","[-_.]*"),"")&IF(B5="|",B5&SUBSTITUTE(A5," ","[-_.]*"),"")
	Row 1 Value	|Reward|Account|Gift[-_.]*card|Voucher|Party[-_.]*planner

Column G	Row 1 Formula	="("&MID(C1,2,1000)&")[-_.]*("&MID(F1,2,1000)&")"
	Row 1 Value	(Amazon|Burger[-_.]*king|Macys|Macy's|Outback[-_.]*steakhouse)[-_.]*(Reward|Account|Gift[-_.]*card|Voucher|Party[-_.]*planner)

The Filter looks like this:

Header	Contains	RegEx	(Amazon|Burger[-_.]*king|Macys|Macy's|Outback[-_.]*steakhouse)[-_.]*(Reward|Account|Gift[-_.]*card|Voucher|Party[-_.]*planner)

5. Another area where you can really optimize your use of Regular Expressions as filters with an MS-Excel spreadsheet is to simply collect the phrases that spammers use with the separators and use them in a spam filter.
    - Spammers use their well-worn phrases (single-word and multiple word phrases) and sometimes use separators to entice us to develop simple filters that might miss their creative iterations.

    - So stay ahead of the game and when you see enough of those emails with certain phrases show, then add their new creations to the filter(s).

        - For Example phrases like:
            - weight loss secret
            - Water Filtration System
            - Virtual reality [anything]
            - Hearing loss reversed
            - Hearing loss treatment
            - Toe Nail Infection
            - Wife walked in
            - Wife out of control
            - Girlfriend walked in
            - Etc.
- Plug these into a simple spreadsheet (Columns A, B, C)
    - Where Colum A contains the text part of the phrase (simple and complex)
    - Column B contains a pipe separator character if Column A is non-blank
    - Column C contains a concatenated Regular Expression filter from the separate Rows 1 through n, where n is your maximum

- Again, I'll use 5 rows as an example. I've expanded this up to 70 rows for my own use.
- Use the same procedure as the compound element test, but only Columns A, B & C and past the results from Column C into the filter
    - NOTE: if it is the first instance of a new filter, you need to enclose the results in parentheses **and remove the leading pipe |. Remember this!!!!**

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 8

o For all subsequent pastes to this filter, go to the end of the filter string, backspace once to position the cursor to the left of the closing right parenthesis ")", then past the resultant string into the filter.

Column A	Row 1	Brain enhancer
	Row 2	(Virtual reality|vr) (goggles|box|glasses|headset)
	Row 3	(weight|Hearing) loss (treatment|control)
	Row 4	Toe nail infection
	Row 5	(wife|girlfriend) (walked in|out of control)
	Etc.	

Column B	Row 1	=IF(ROW(A1)=1,"|",IF(A2=A1,"*","|"))
	Row 2	=IF(ROW(A2)=1,"|",IF(A3=A2,"*","|"))
	Row 3	=IF(ROW(A3)=1,"|",IF(A4=A3,"*","|"))
	Row 4	=IF(ROW(A4)=1,"|",IF(A5=A4,"*","|"))
	Row 5	=IF(ROW(A5)=1,"|",IF(A6=A5,"*","|"))
	Etc.	

Column C	Row 1 Formula	=IF(B1="|",B1&SUBSTITUTE(A1," ","[-_.]*"),"")&IF(B2="|",B2&SUBSTITUTE(A2," ","[-_.]*"),"")&IF(B3="|",B3&SUBSTITUTE(A3," ","[-_.]*"),"")&IF(B4="|",B4&SUBSTITUTE(A4," ","[-_.]*"),"")&IF(B5="|",B5&SUBSTITUTE(A5," ","[-_.]*"),"")
	Row 1 Value (filter)	|Brain[-_.]*enhancer|(Virtual[-_.]*reality|vr)[-_.]*(goggles|box|glasses|headset)|(weight|Hearing)[-_.]*loss[-_.]*(treatment|control)|Toe[-_.]*nail[-_.]*infection|(wife|girlfriend)[-_.]*(walked[-_.]*in|out[-_.]*of[-_.]*control)

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 9

The Filter looks like this:

NOTE: if it is the first instance of a new filter, you need to enclose the results in parentheses **and remove the leading pipe |. Remember this!!!!**

Header	Contains	RegEx	(Brain[-_.]*enhancer|(Virtual[-_.]*reality|vr)[-_.]*(goggles|box|glasses|headset)|(weight|Hearing)[-_.]*loss[-_.]*(treatment|control)|Toe[-_.]*nail[-_.]*infection|(wife|girlfriend)[-_.]*(walked[-_.]*in|out[-_.]*of[-_.]*control))

NOTE: If this is not the first instance of a new filter, go to the end of the filter string, backspace once to position the cursor to the left of the closing right parenthesis ")", then paste the string into the filter.

Header	Contains	RegEx	(Hire Offshore Developer|Red pottery pot terracotta|TransUnion, Equifax, and Experian|Compare Medicare Plans|Research studies may offer payment|New Fat Burner|CVS by Storm|Penny Pot Stock|Roof is covered|Revolutionary Non-Stick|Scratch Resistant Pan|Medicare Enrollment Period|Election Sale|Become a Wall Street Journal Member|Dear in Christ|Chronic Constipation|Best-boost for you|your loving gun|Borrow from a trusted|crafty psychological trick|3 Things Jesus Said|About How to Cure Disease)|(Brain[-_.]*enhancer|(Virtual[-_.]*reality|vr)[-_.]*(goggles|box|glasses|headset)|(weight|Hearing)[-_.]*loss[-_.]*(treatment|control)|Toe[-_.]*nail[-_.]*infection|(wife|girlfriend)[-_.]*(walked[-_.]*in|out[-_.]*of[-_.]*control))

Short Primer #02 on Using Regular Expressions (RegEx) with MailWasherPro Filters and Rules – page 10

- **Checking for singular versus plurals**
  - For simple plurals (ending in s), place the "s" at the end of the root word in the RegEx followed by a question mark (?).

Header	Contains	RegEx	(stocks?|bonds?|stores?)
Header	Contains	RegEx	(stock|bond|store)s?

  - The above expression will check for both singular and plural.
  - The question mark simply checks for the string with or without the character preceding the question mark
  - The second example optimizes the expression

  - For complex plurals that use other endings, you would use a more the full form of the plural

Header	Contains	RegEx	(memories|specialties|wineries|refineries|bakeries)
Header	Contains	RegEx	(memor|specialt|winer|refiner|baker)ies

- **Checking for compound or multiple "things"**
  - This principle for the construction of complex plurals should lead you to understand that you can construct expressions to search for other endings for not only singular and plural forms of root words, but also for words that have the same stem but that are not related:

Header	Contains	RegEx	restor(es?|ing|ed|er|ative)
Header	Contains	RegEx	cop(y|ier|ying|ied|per|ulate|e|ed|ing|tic)
Header	Contains	RegEx	(con|de|ob)(struction|serve|figure)

  - The main thing to remember is that you are fighting against a clever adversary who is trying to complicate your life so exercise your creativity in finding ways to anticipate what terminology they might try to use and build defenses against it.
  - Sometimes you will create combinations that are nonsense (such as obfigure), but if it helps you build regular expressions that catch a lot of real stuff defeat that adversary's garbage, then that's OK.