Using resources with regex in memoQ

… all the power & NONE of the syntax!

(Winter Translation Forum '23)

Kevin Lossner I December 1, 2023 I WTF workshop

Today’s topics

1 **Regex in memoQ**

- What good is it?
- Where can it be used?
- Syntax vs. application

2 **Getting started: the RA library**

- Regex Assistant “anatomy”
- Navigation & features
- Live examples (“demoQ”)

3 **More resources**

- Web tools
- memoQuickies Resource Camp
- People & places

4 **Open Q&A**

- Questions

Why are regular expressions useful?

A *description language* for handling *patterned* text

C_2H_5OH 3.520 g £3.50 €24.2bn USD 24,000

25 November 2020 / June 23rd 2015 2:15 p.m., 2:15 PM, 14h15, 14:15

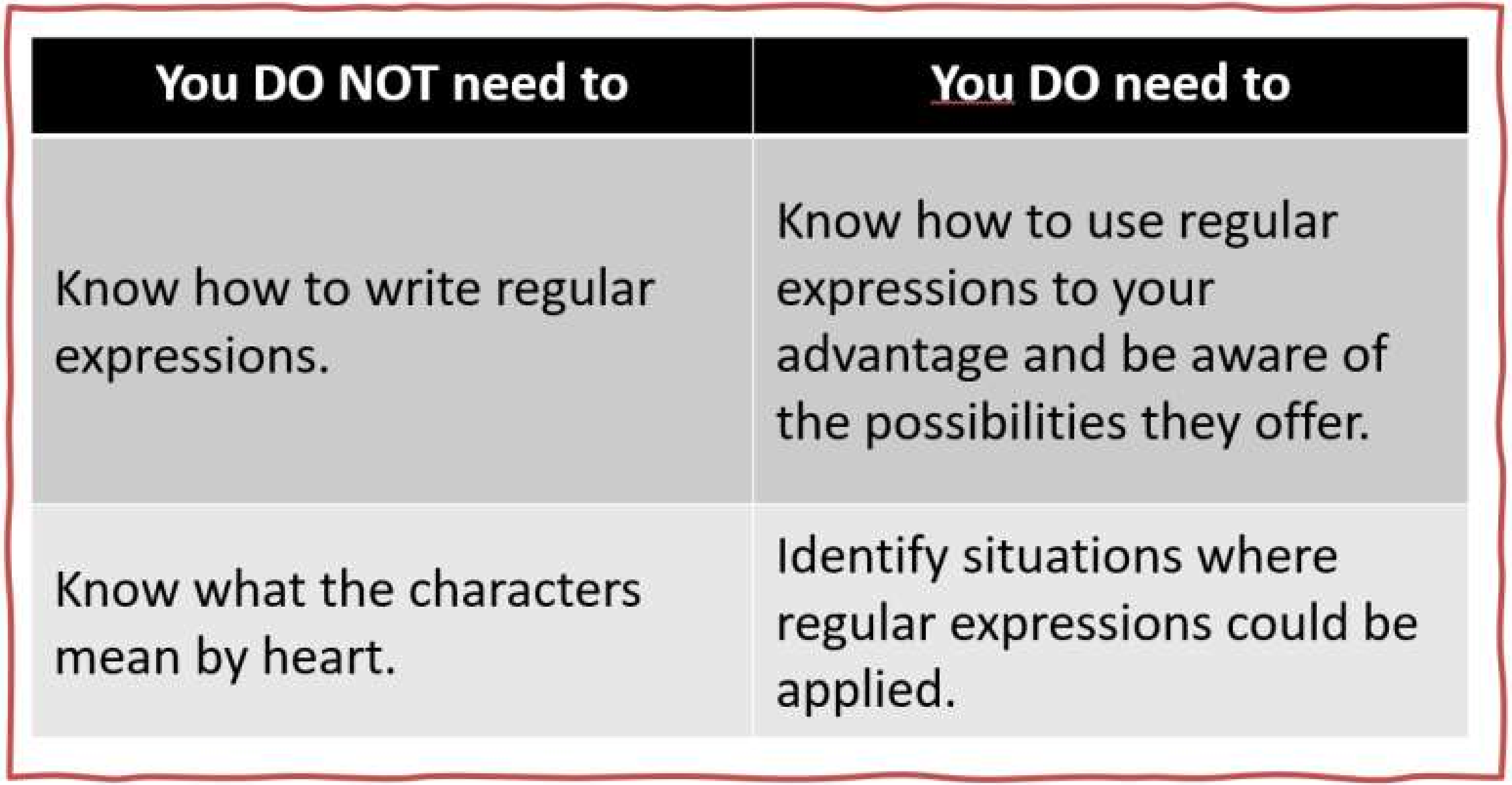
Shiau v. U.S. Dep't of Agric., 895 F.2d 1410 (2d Cir. 1989).

In memoQ

- Better segmentation
- New or better import filters
- “Tagging” text to protect it
- Placeable translation suggestions

- Filtering source + target
- Find & replace
  >> dialog
  >> external automation tool
- QA profile checks
- Easy-to-use **library**

Priorities for problem-solving



You DO NOT need to	You DO need to
Know how to write regular expressions.	Know how to use regular expressions to your advantage and be aware of the possibilities they offer.
Know what the characters mean by heart.	Identify situations where regular expressions could be applied.

Priorities for planning & expert assistance

Describe the "problem" to be solved clearly

- What kind of information and for what purpose?
- Comprehensive examples, including exceptions and errors
- Solution use (find? fix? source/target? what feature fits best?)
- Is there a need for precision (date validation, etc.)?

ChatGPT won't help!!!

Example: admission price info in *target* text. Correct all the myriad formats in the DeepL MT spew to €X.XX

€ X.XX €X,XX EUR X,XX X.XX€ X,XX euros X.XX EUR

- Want "screening filters", find&replace regexes, auto-translation

memoQ modules with regex

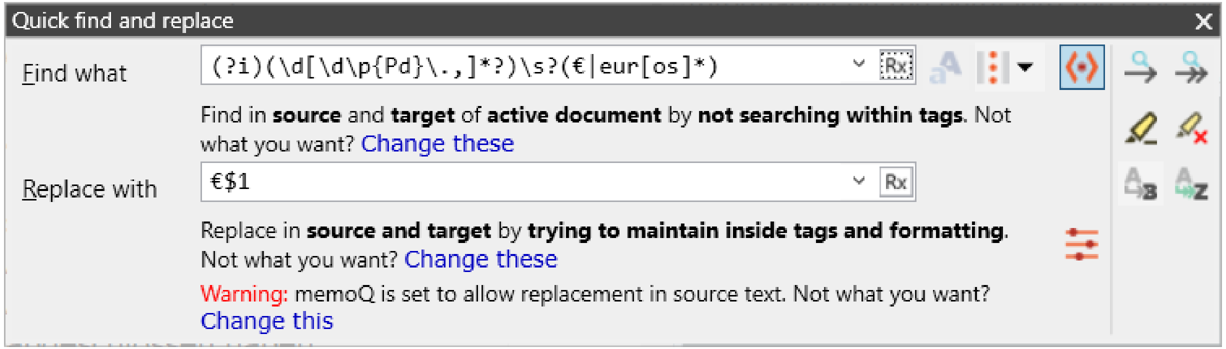
- Segmentation rules
- Filters: Regex Text, ZIP, etc.
- Regex Tagger (can be saved for cascading filters)
- Auto-translation rules
- QA profiles

PLUGANDPLAY

memoQ modules with regex

Working grid (filter fields)

Find & Replace dialogs



Quick find and replace

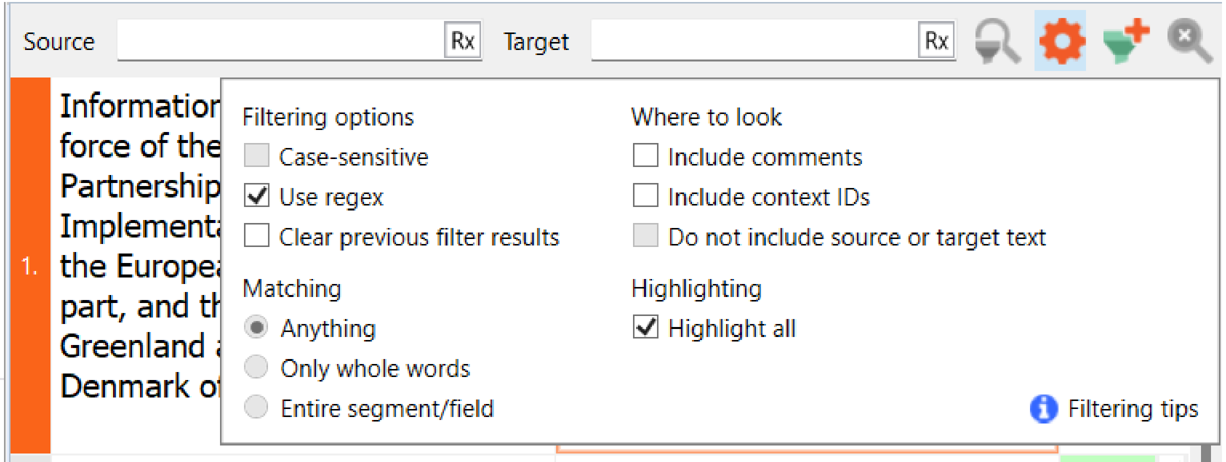
Find what `(?i)(\d[\d\p{Pd}\.,]*?)\s?(€|eur[os]*)` Rx

Find in **source** and **target** of **active document** by **not searching within tags**. Not what you want? Change these

Replace with `€$1` Rx

Replace in **source and target** by **trying to maintain inside tags and formatting**. Not what you want? Change these

Warning: memoQ is set to allow replacement in source text. Not what you want? Change this



Source Rx Target Rx

1. Information force of the Partnership Implementa the Europe part, and th Greenland a Denmark o

Filtering options

□ Case-sensitive

☑ Use regex

□ Clear previous filter results

Matching

● Anything

○ Only whole words

○ Entire segment/field

Where to look

□ Include comments

□ Include context IDs

□ Do not include source or target text

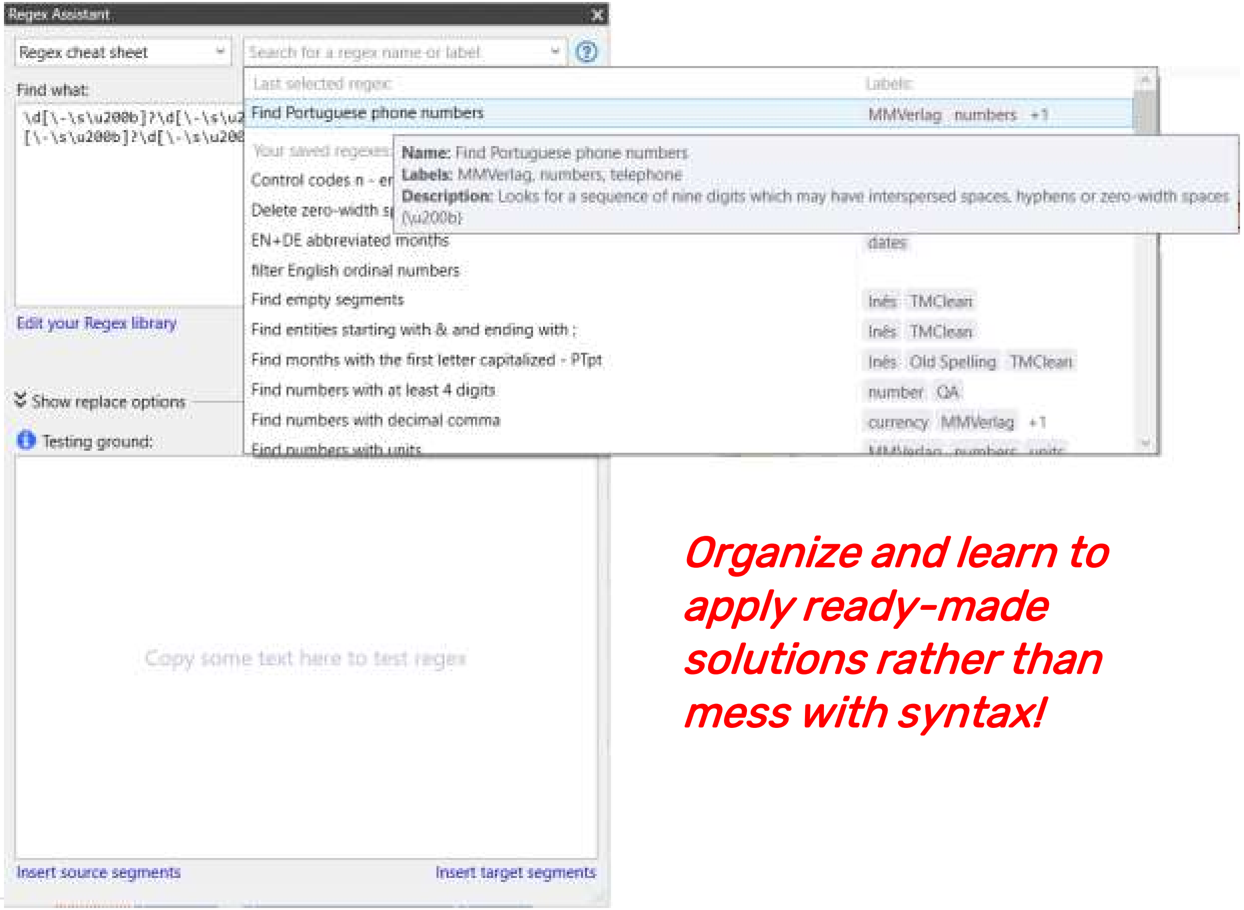
Highlighting

☑ Highlight all

Filtering tips

Regex Assistant

A built-in library for managing regex info with plain language names, labels and descriptions



Regex Assistant

Regex cheat sheet

Search for a regex name or label

Find what:

```
\d[\-\s\u200b]?\d[\-\s\u2
[\-\s\u200b]?\d[\-\s\u200
```

Last selected regex	Labels
Find Portuguese phone numbers	MMVerlag numbers +1
Your saved regexes	
Control codes n - er	
Delete zero-width sp	
EN+DE abbreviated months	dates
filter English ordinal numbers	
Find empty segments	Inês TMClean
Find entities starting with & and ending with ;	Inês TMClean
Find months with the first letter capitalized - PTpt	Inês Old Spelling TMClean
Find numbers with at least 4 digits	number QA
Find numbers with decimal comma	currency MMVerlag +1
Find numbers with units	MMVerlag numbers units

Name: Find Portuguese phone numbers
Labels: MMVerlag, numbers, telephone
Description: Looks for a sequence of nine digits which may have interspersed spaces, hyphens or zero-width spaces (\u200b)

Edit your Regex library

Show replace options

Testing ground:

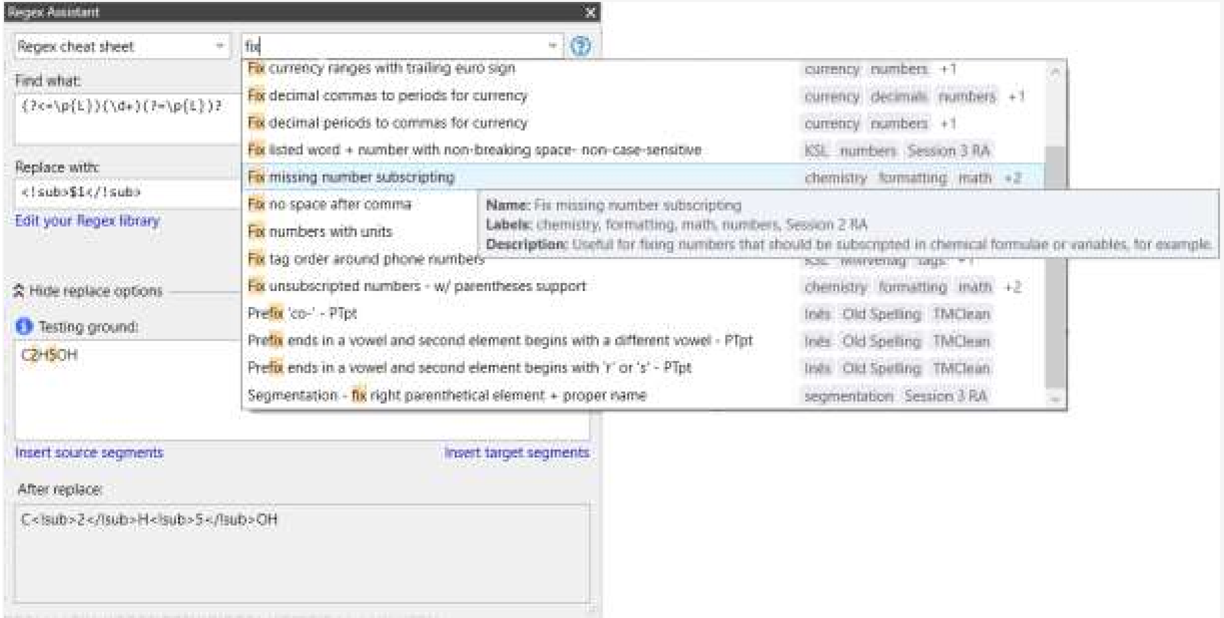
Copy some text here to test regex

Insert source segments

Insert target segments

Organize and learn to apply ready-made solutions rather than mess with syntax!

The Regex Assistant



Regex Assistant

Regex cheat sheet

fix

Find what:

```
(?<=\p{L})(\d+)(?=\p{L})?
```

Replace with:

```
<!sub>$1</!sub>
```

Edit your Regex library

Hide replace options

Testing ground:

C2H5OH

Insert source segments

Insert target segments

After replace:

C<!sub>2</!sub>H<!sub>5</!sub>OH

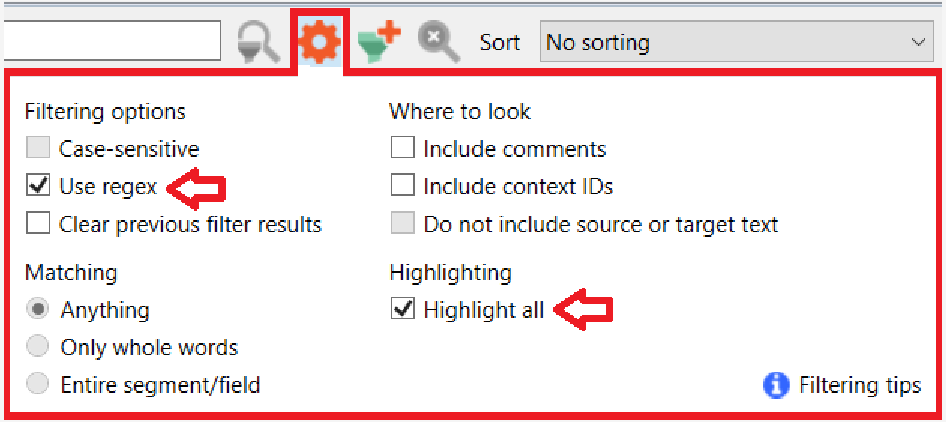
Fix currency ranges with trailing euro sign	currency numbers +1
Fix decimal commas to periods for currency	currency decimals numbers +1
Fix decimal periods to commas for currency	currency numbers +1
Fix listed word + number with non-breaking space- non-case-sensitive	KSL numbers Session 3 RA
Fix missing number subscripting	chemistry formatting math +2
Fix no space after comma	
Fix numbers with units	
Fix tag order around phone numbers	KSL intervening tags +1
Fix unsubscripted numbers - w/ parentheses support	chemistry formatting math +2
Prefix 'co-' - PTpt	Inês Old Spelling TMClean
Prefix ends in a vowel and second element begins with a different vowel - PTpt	Inês Old Spelling TMClean
Prefix ends in a vowel and second element begins with 'r' or 's' - PTpt	Inês Old Spelling TMClean
Segmentation - fix right parenthetical element + proper name	segmentation Session 3 RA

Name: Fix missing number subscripting
Labels: chemistry, formatting, math, numbers, Session 2 RA
Description: Useful for fixing numbers that should be subscripted in chemical formulae or variables, for example.

Access & Features

The Regex Assistant for filtering in the Translation & Editing Grid

1 **Settings > Use regex**



Filtering options

- ☐ Case-sensitive
- ☑ Use regex
- ☐ Clear previous filter results

Where to look

- ☐ Include comments
- ☐ Include context IDs
- ☐ Do not include source or target text

Matching

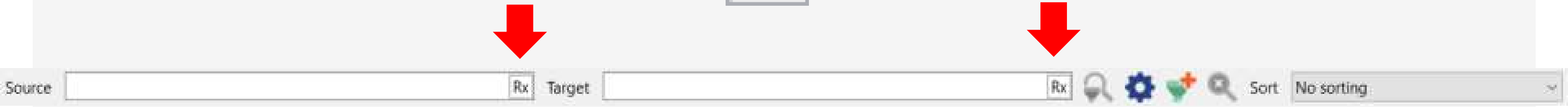
- ● Anything
- ○ Only whole words
- ○ Entire segment/field

Highlighting

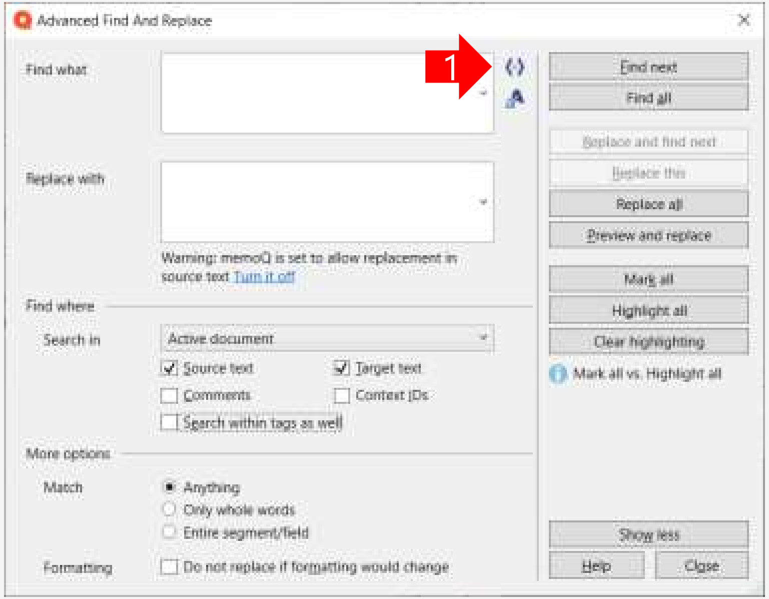
- ☑ Highlight all

ℹ Filtering tips

2 **Click the Regex Assistant icon!** Rx



The Regex Assistant for the Find & Replace dialogs (1 of 2)



Find what

Replace with

Warning: memoQ is set to allow replacement in source text Turn it off

Find where

Search in: Active document

☑ Source text ☑ Target text

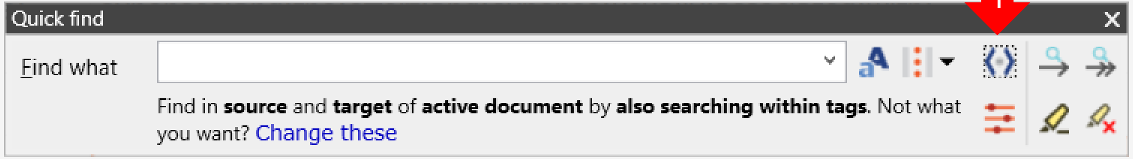
□ Comments □ Context IDs

□ Search within tags as well

More options

Match: ● Anything ○ Only whole words ○ Entire segment/field

Formatting: □ Do not replace if formatting would change



Find what

Find in **source** and **target** of **active document** by **also searching within tags**. Not what you want? Change these

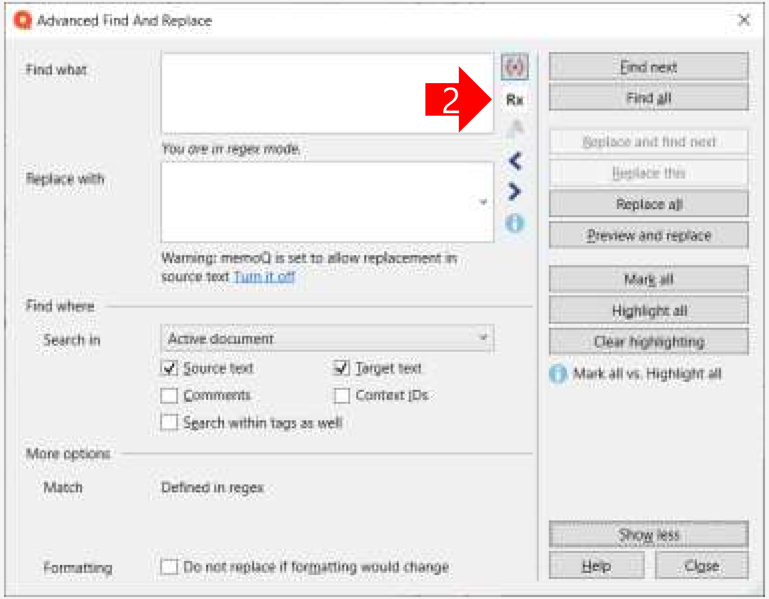
1 **Click the blue icon to turn on the use of regular expressions. It will change**

... and the Regex Assistant icon will appear in the dialog.

Rx

The Regex Assistant for the Find & Replace dialogs (2 of 2)

2 **Click the Regex Assistant icon** Rx



Find what

You are in regex mode.

Replace with

Warning: memoQ is set to allow replacement in source text Turn it off

Find where

Search in: Active document

☑ Source text ☑ Target text

□ Comments □ Context IDs

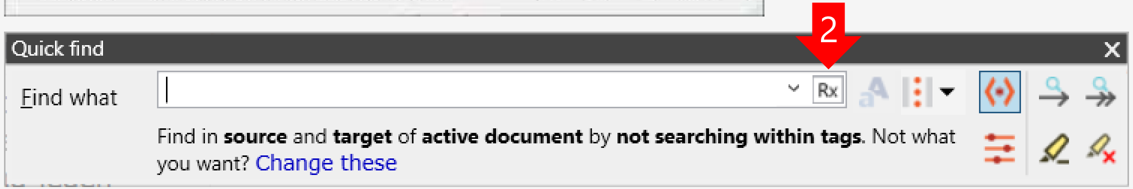
□ Search within tags as well

Mark all vs. Highlight all

More options

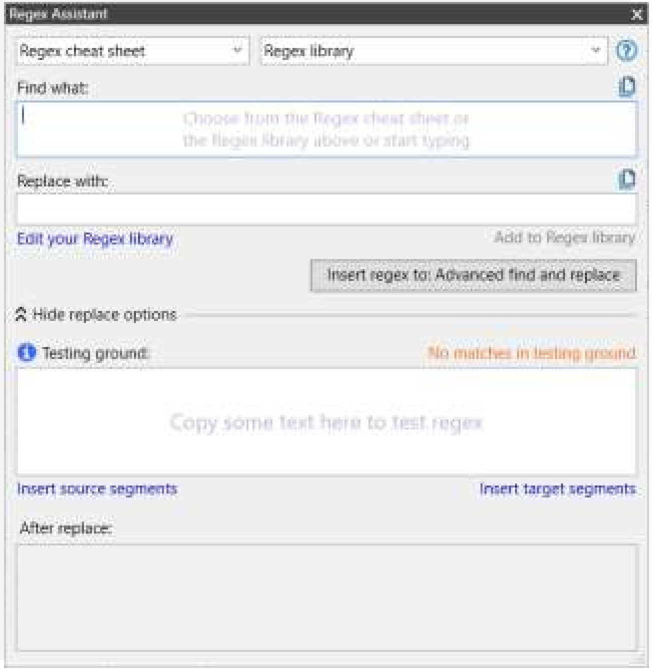
Match: Defined in regex

Formatting: □ Do not replace if formatting would change



Find what

Find in **source** and **target** of **active document** by **not searching within tags**. Not what you want? Change these



Find what:

Choose from the Regex cheat sheet or the Regex library above or start typing

Replace with:

Edit your Regex library

Add to Regex library

Hide replace options

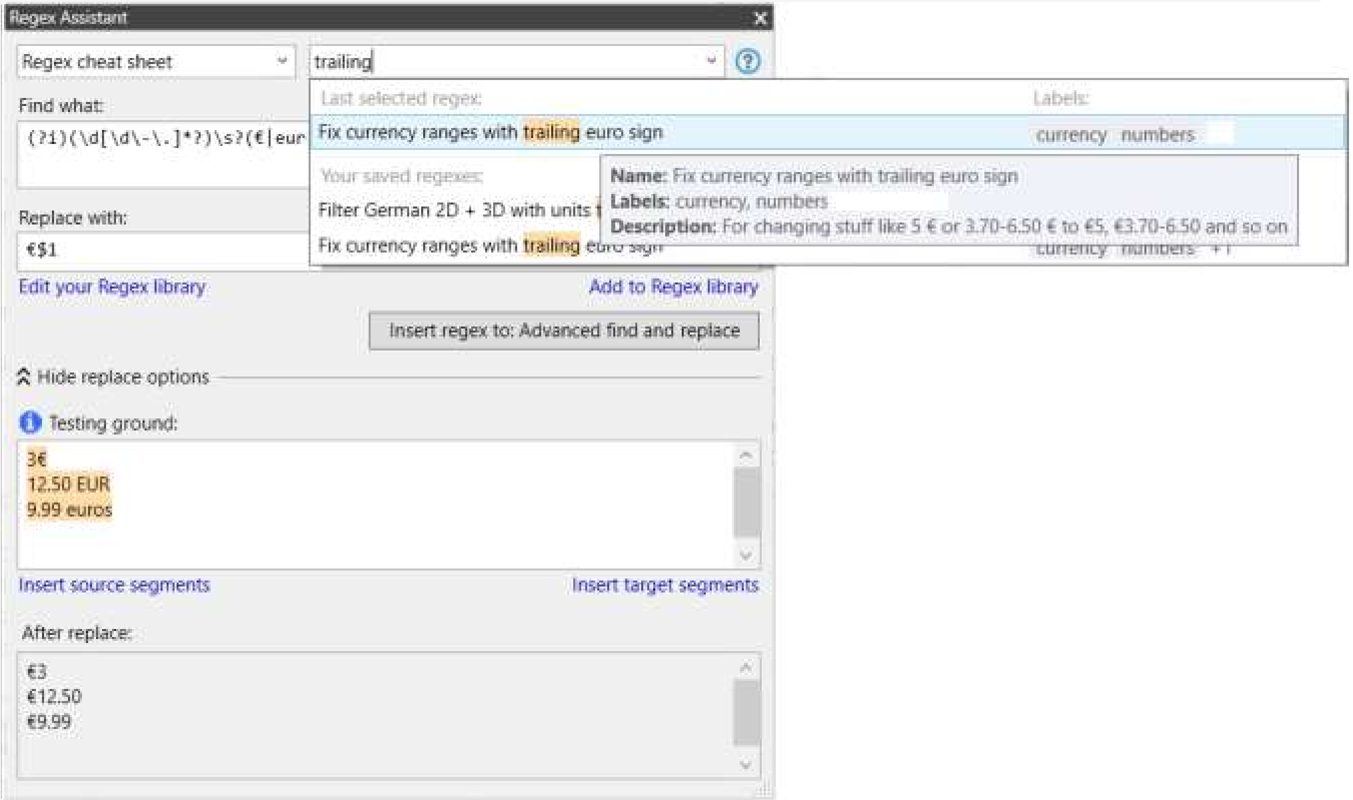
Testing ground:

No matches in testing ground

Copy some text here to test regex

After replace:

“Anatomy” of the memoQ Regex Assistant



Regex cheat sheet

trailing

Fix currency ranges with trailing euro sign — currency numbers

Filter German 2D + 3D with units

Fix currency ranges with trailing euro sign — currency numbers +1

Name: Fix currency ranges with trailing euro sign
Labels: currency, numbers
Description: For changing stuff like 5 € or 3.70-6.50 € to €5, €3.70-6.50 and so on

Find what:

```
(?i)(\d[\d\-\.]*?)\s?(€|eur
```

Replace with:

```
€$1
```

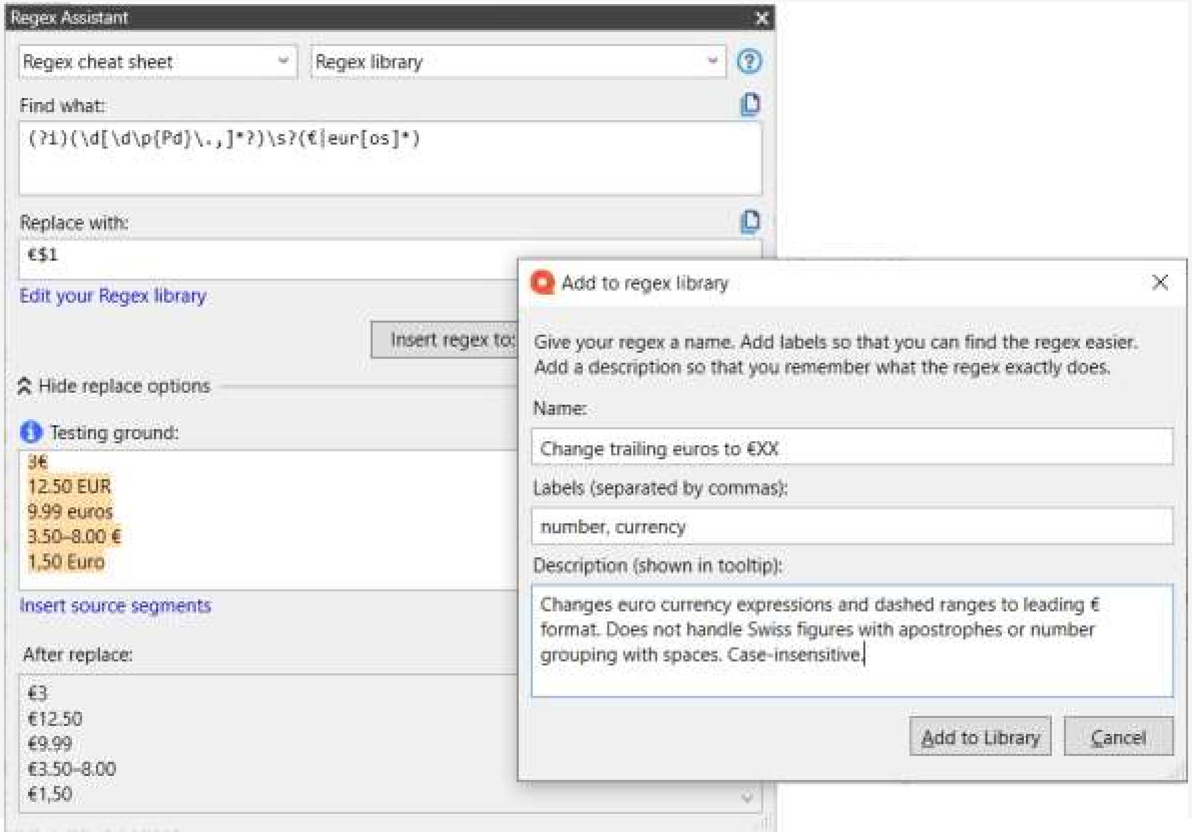
Testing ground:

```
3€
12.50 EUR
9.99 euros
```

After replace:

```
€3
€12.50
€9.99
```

“Anatomy” of the memoQ Regex Assistant



Find what:

```
(?i)(\d[\d\p{Pd}\.,]*?)\s?(€|eur[os]*)
```

Replace with:

```
€$1
```

Testing ground:

```
3€
12.50 EUR
9.99 euros
3.50–8.00 €
1,50 Euro
```

After replace:

```
€3
€12.50
€9.99
€3.50–8.00
€1,50
```

Give your regex a name. Add labels so that you can find the regex easier. Add a description so that you remember what the regex exactly does.

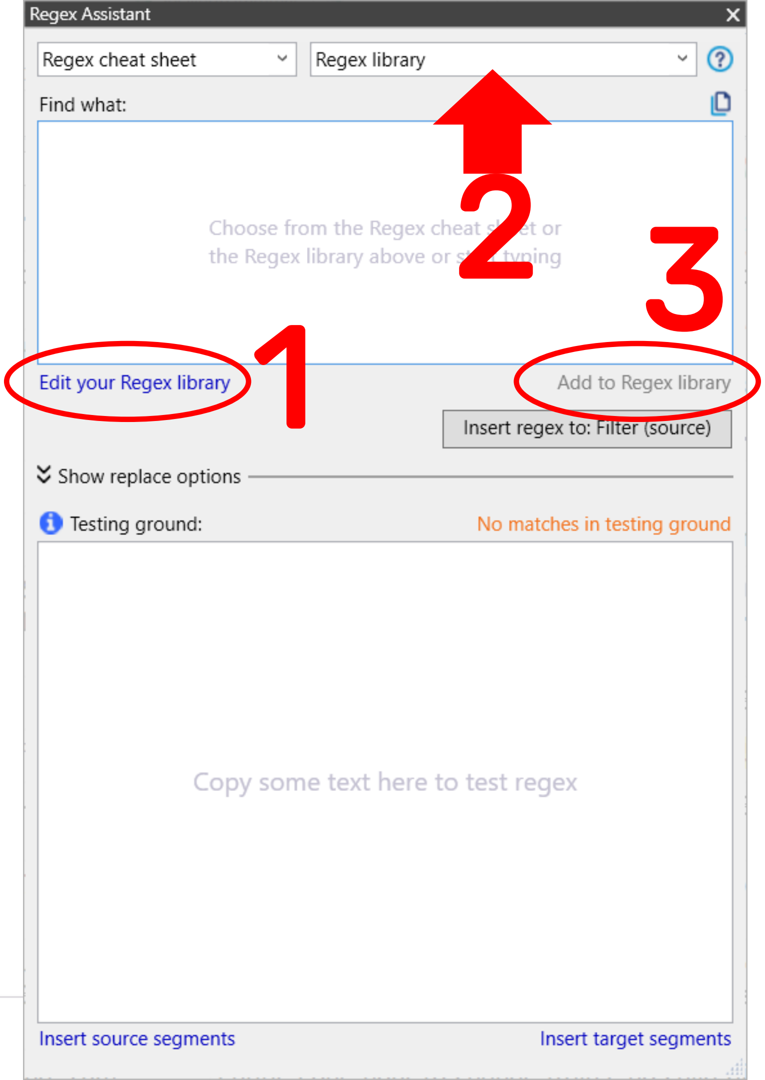
Name: Change trailing euros to €XX

Labels (separated by commas): number, currency

Description (shown in tooltip): Changes euro currency expressions and dashed ranges to leading € format. Does not handle Swiss figures with apostrophes or number grouping with spaces. Case-insensitive.

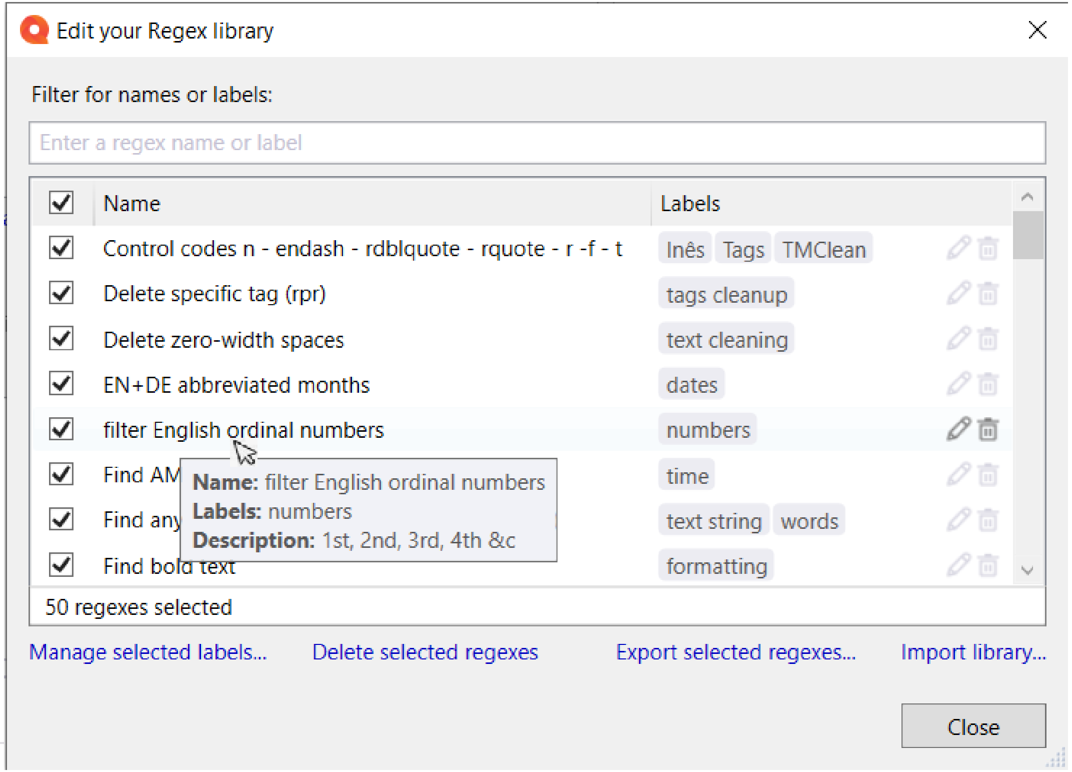
Exchange, edits & adding

1. To import or export library exchange files (XML) or change stored expressions, click "*Edit your Regex library*".
2. Use the "*Regex library*" search field to find the stored expression you need.
3. Select/copy interesting regexes and "*Add to Regex library*".



Exchange, edits & access

The library’s editing dialog is where you select expressions to export and share, import regex collections from others, edit names, descriptions and labels, etc.



Edit your Regex library

Filter for names or labels:

Enter a regex name or label

☑	Name	Labels
☑	Control codes n - endash - rdblquote - rquote - r -f - t	Inês Tags TMClean
☑	Delete specific tag (rpr)	tags cleanup
☑	Delete zero-width spaces	text cleaning
☑	EN+DE abbreviated months	dates
☑	filter English ordinal numbers	numbers
☑	Find AM	time
☑	Find any	text string words
☑	Find bold text	formatting

Name: filter English ordinal numbers
Labels: numbers
Description: 1st, 2nd, 3rd, 4th &c

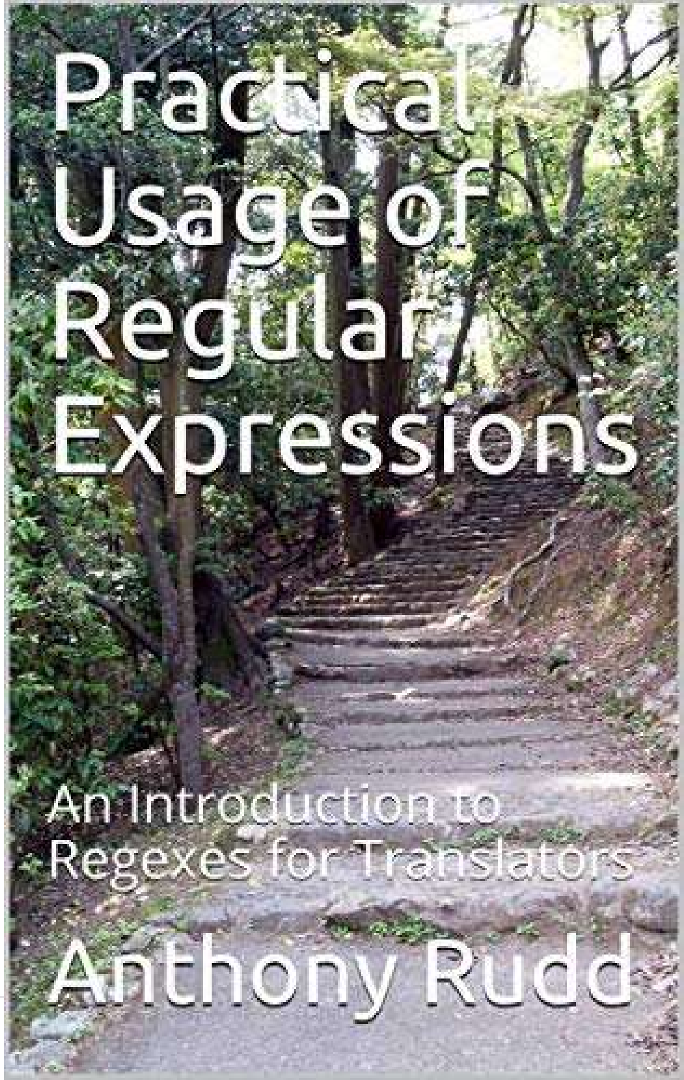
50 regexes selected

DEMOQ

For aspiring syntax nerds and other masochists….

#2 Recommendation: **Anthony Rudd**

Practical Usage of Regular Expressions – An Introduction to Regexes for Translators

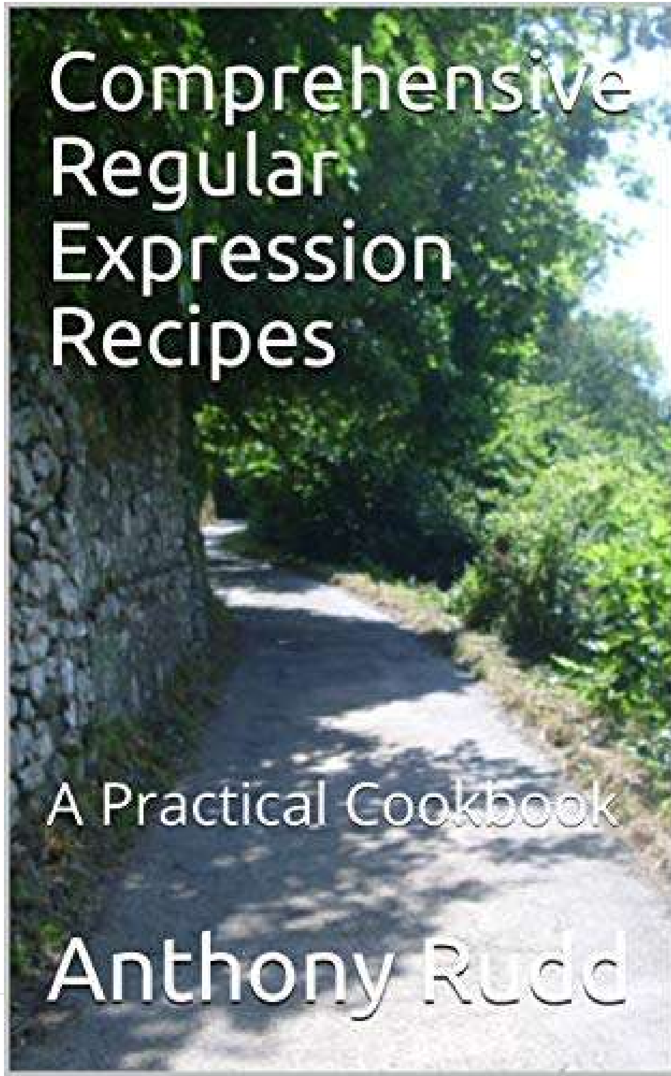


Practical Usage of Regular Expressions
An Introduction to Regexes for Translators
Anthony Rudd

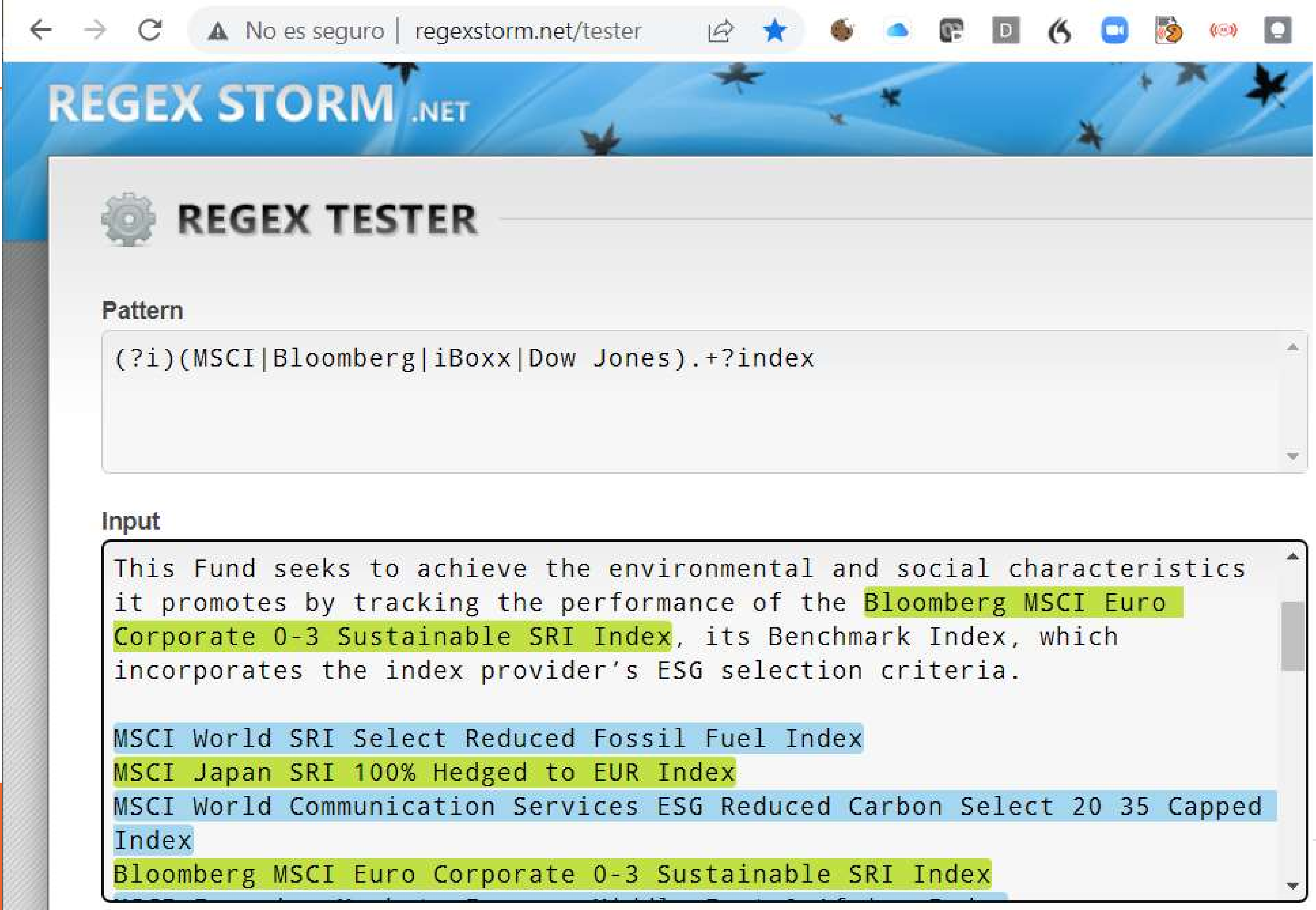
Probably more useful:

#1 Recommendation: **Anthony Rudd**
Comprehensive Regular Expression Recipes – A Practical Cookbook

Nearly 100 pages of "solved" common translation text problems!



Web resources



Pattern

```
(?i)(MSCI|Bloomberg|iBoxx|Dow Jones).+?index
```

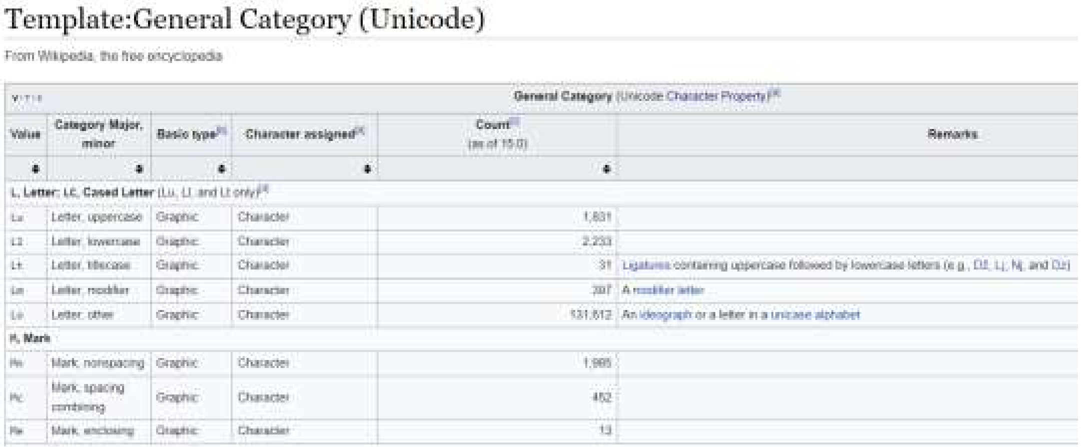
Input

This Fund seeks to achieve the environmental and social characteristics it promotes by tracking the performance of the Bloomberg MSCI Euro Corporate 0-3 Sustainable SRI Index, its Benchmark Index, which incorporates the index provider's ESG selection criteria.

MSCI World SRI Select Reduced Fossil Fuel Index
MSCI Japan SRI 100% Hedged to EUR Index
MSCI World Communication Services ESG Reduced Carbon Select 20 35 Capped Index
Bloomberg MSCI Euro Corporate 0-3 Sustainable SRI Index

Web resources

More reference information



Template:General Category (Unicode)

From Wikipedia, the free encyclopedia

V · T · E

General Category (Unicode Character Property)

Value	Category Major, minor	Basic type	Character assigned	Count (as of 15.0)	Remarks
L, Letter; LC, Cased Letter (Lu, Ll, and Lt only)					
Lu	Letter, uppercase	Graphic	Character	1,831	
Ll	Letter, lowercase	Graphic	Character	2,233	
Lt	Letter, titlecase	Graphic	Character	31	Ligatures containing uppercase followed by lowercase letters (e.g., ǅ, ǈ, ǋ, and ǲ)
Lm	Letter, modifier	Graphic	Character	397	A modifier letter
Lo	Letter, other	Graphic	Character	131,612	An ideograph or a letter in a unicase alphabet
M, Mark					
Mn	Mark, nonspacing	Graphic	Character	1,985	
Mc	Mark, spacing combining	Graphic	Character	452	
Me	Mark, enclosing	Graphic	Character	13	

General categories in Unicode allow more concise expressions w/ greater scope: https://en.wikipedia.org/wiki/Template:General_Category_(Unicode)

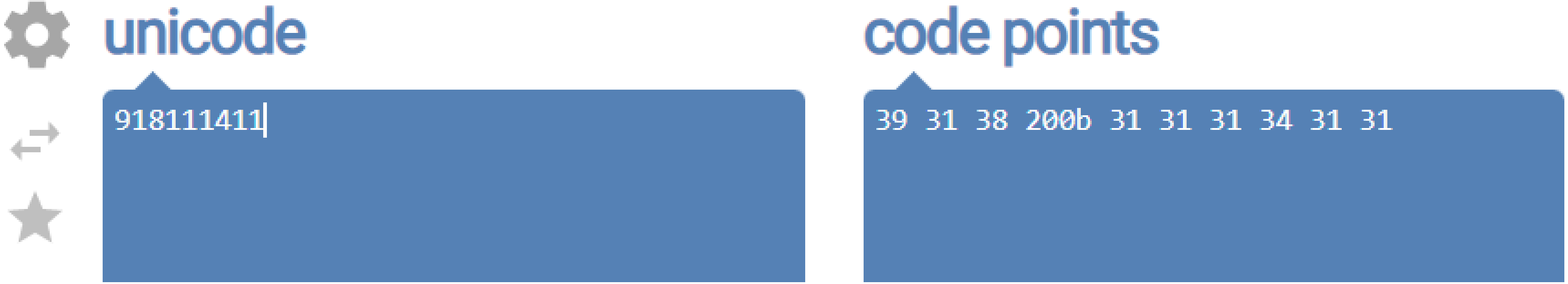
For example:

`\p{L}` instead of `[a-zA-Z]`, also covering Cyrillic, etc.

`\p{Pi}` and `\p{Pf}` for 12 initial and 10 final quotation marks respectively (excluding ASCII straight quotes)

More web tools

https://onlineunicodetools.com/convert-unicode-to-code-points



unicode

```
918111411
```

code points

```
39 31 38 200b 31 31 31 34 31 31
```

Web resources

More reference information (template exemples for regex)

Link to Marek’s Memsource solution (Find&Replace tool example):
http://wasaty.pl/blog/2018/01/30/memsource-files-in-memoq/

Link to memoQ Help on the Find & Replace tool:
https://docs.memoq.com/current/en/Places/edit-template-find-and-replace.html

Link to my review of Marek’s template book
https://www.translationtribulations.com/2022/06/local-memoQ-templates-guide.html

Web resources

memoQ Help: https://docs.memoq.com/current/en/Places/regex-assistant.html

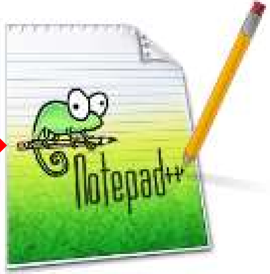
Lists a few places in memoQ with regex capabilities that I have not talked about, with links to the relevant Help pages:

- memoQ's ZIP filter
- The MSG filter for e-mail
- The Length tab on QA profiles

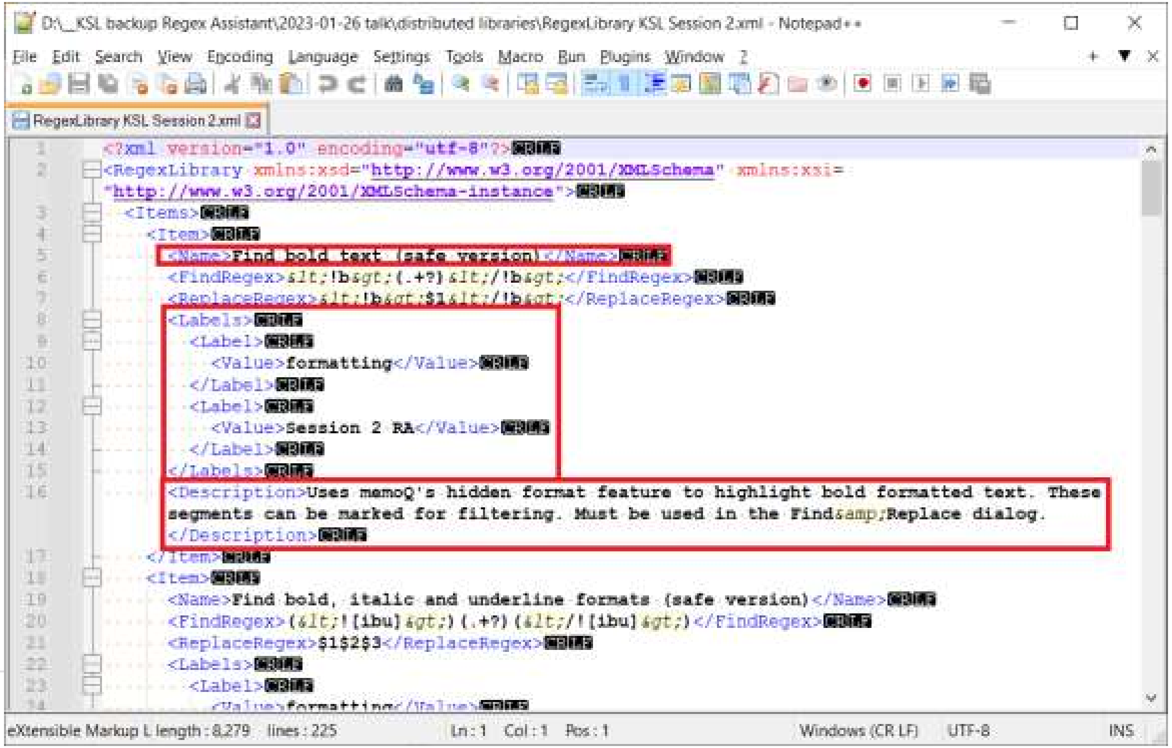
Translating Regex Assistant libraries!

A configured filter has been provided for those who want to translate the names, labels and descriptions of expressions into another language.

Filter provided



XMLConverter#Regex Assistant library XML export.mqres


D:_KSL backup Regex Assistant\2023-01-26 talk\distributed libraries\RegexLibrary KSL Session 2.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
RegexLibrary KSL Session 2.xml

```
<?xml version="1.0" encoding="utf-8"?>
<RegexLibrary xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance">
  <Items>
    <Item>
      <Name>Find bold text (safe version)</Name>
      <FindRegex>&lt;!b&gt;(.+?)&lt;/!b&gt;</FindRegex>
      <ReplaceRegex>&lt;!b&gt;$1&lt;/!b&gt;</ReplaceRegex>
      <Labels>
        <Label>
          <Value>formatting</Value>
        </Label>
        <Label>
          <Value>Session 2 RA</Value>
        </Label>
      </Labels>
      <Description>Uses memoQ's hidden format feature to highlight bold formatted text. These
segments can be marked for filtering. Must be used in the Find&amp;Replace dialog.
</Description>
    </Item>
    <Item>
      <Name>Find bold, italic and underline formats (safe version)</Name>
      <FindRegex>(&lt;![ibu]&gt;)(.+?)(&lt;/![ibu]&gt;)</FindRegex>
      <ReplaceRegex>$1$2$3</ReplaceRegex>
      <Labels>
        <Label>
          <Value>formatting</Value>
```

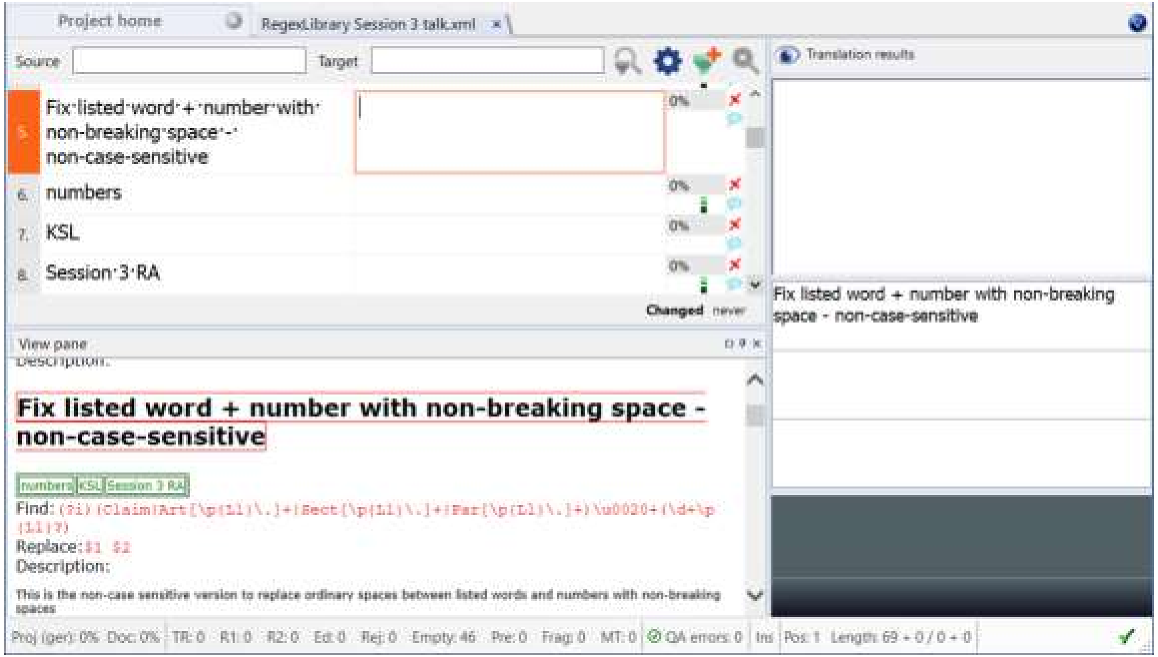
eXtensible Markup L length : 8,279 lines : 225 Ln : 1 Col : 1 Pos : 1 Windows (CR LF) UTF-8 INS

Translating Regex Assistant libraries!

A configured filter has been provided for those who want to translate the names, labels and descriptions of expressions into another language.

Filter provided

XMLConverter#Regex Assistant library XML export.mqres



	Source	Target	
5.	Fix·listed·word·+·number·with·non-breaking·space·-·non-case-sensitive		0%
6.	numbers		0%
7.	KSL		0%
8.	Session·3·RA		0%

Fix listed word + number with non-breaking space - non-case-sensitive

Description:

Fix listed word + number with non-breaking space - non-case-sensitive

numbers KSL Session 3 RA

Find: `(?i)(Claim|Art[\p{Ll}\.]+|Sect[\p{Ll}\.]+|Par[\p{Ll}\.]+)\u0020+(\d+\p{Ll}?)`

Replace: `$1 $2`

Description:

This is the non-case sensitive version to replace ordinary spaces between listed words and numbers with non-breaking spaces

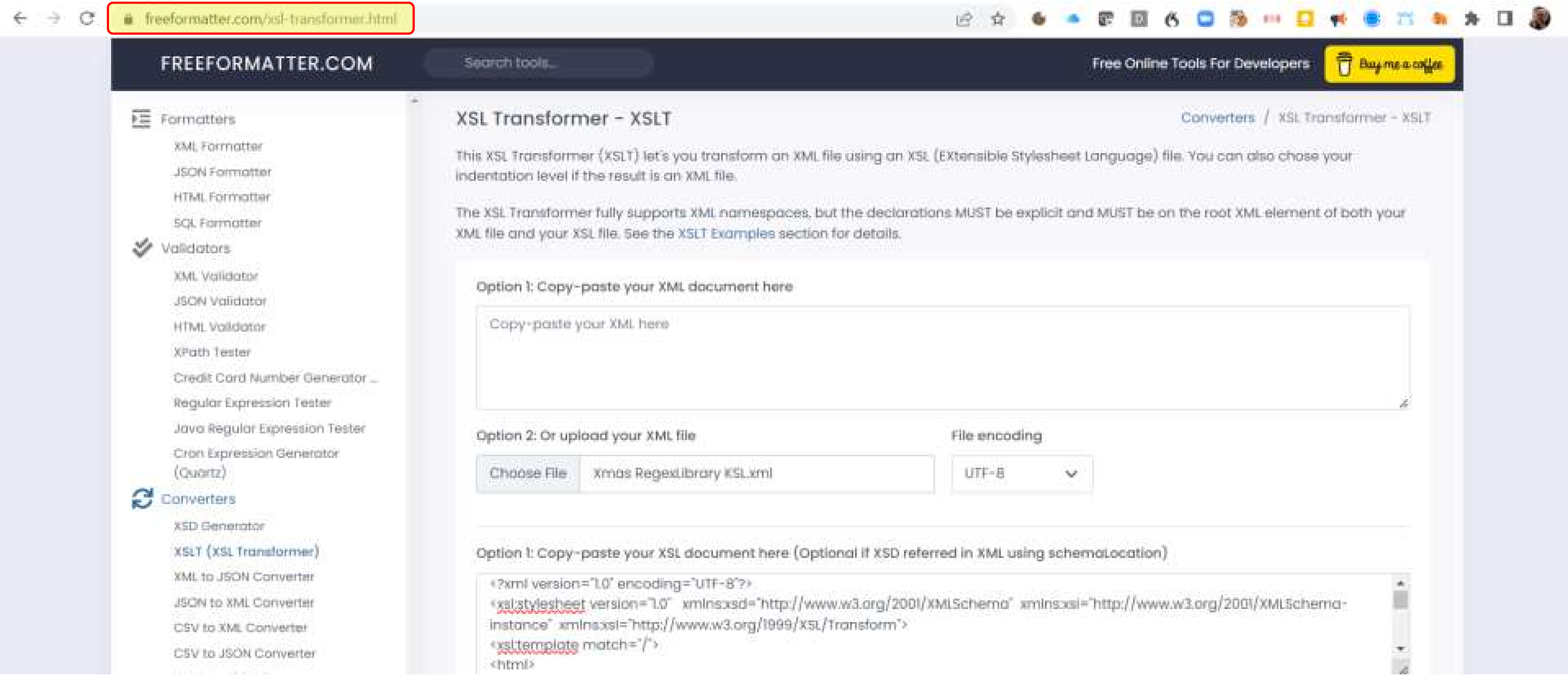
```
Proj (ger): 0%  Doc: 0%  TR: 0  R1: 0  R2: 0  Ed: 0  Rej: 0  Empty: 46  Pre: 0  Frag: 0  MT: 0  QA errors: 0  Ins  Pos: 1  Length: 69 + 0 / 0 + 0
```

Readable output with XSLT (1 of 3)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"   xmlns:xsd="http://www.w3.org/2001/XMLSchema"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/">
<html>
<body>
<xsl:for-each select="/RegexLibrary/Items/Item">
<H3 style="font-family:verdana;font-size:24px;font-style:bold;color:black"><xsl:value-of select="Name"/></H3>
<table border="1" valign="center" style="font-family:verdana;font-size:12px;color:green"><tr align="center">
<xsl:for-each select="Labels/Label">
 <td><xsl:value-of   select="Value"/></td>
</xsl:for-each>
</tr></table>
<span style="font-family:verdana;font-size:16px;font-style:bold;color:black">Find:</span> <span type="FindRegex" style="font-
family:courier;font-size:16px;font-style:bold;color:red"><xsl:value-of select="FindRegex"/></span><br />
<span style="font-family:verdana;font-size:16px;font-style:bold;color:black">Replace:</span> <span type="ReplaceRegex" style="font-
family:courier;font-size:16px;font-style:bold;color:red"><xsl:value-of select="ReplaceRegex"/></span><br />
<span style="font-family:verdana;font-size:16px;font-style:bold;color:black">Description:</span>
<p type="Description" style="font-family:verdana;font-size:12px;color:black">
<xsl:value-of   select="Description"/>
</p>
</xsl:for-each>
</body>
</html>
</xsl:template>
</xsl:stylesheet>
```

This XSL script can also be used for a preview if translating exported Regex Assistant libraries (XML).

Readable output with XSLT (2 of 3)



freeformatter.com/xsl-transformer.html

FREEFORMATTER.COM

Search tools...

Free Online Tools For Developers

Buy me a coffee

XSL Transformer - XSLT

Converters / XSL Transformer - XSLT

This XSL Transformer (XSLT) let's you transform an XML file using an XSL (EXtensible Stylesheet Language) file. You can also chose your indentation level if the result is an XML file.

The XSL Transformer fully supports XML namespaces, but the declarations MUST be explicit and MUST be on the root XML element of both your XML file and your XSL file. See the XSLT Examples section for details.

Option 1: Copy-paste your XML document here

Copy-paste your XML here

Option 2: Or upload your XML file

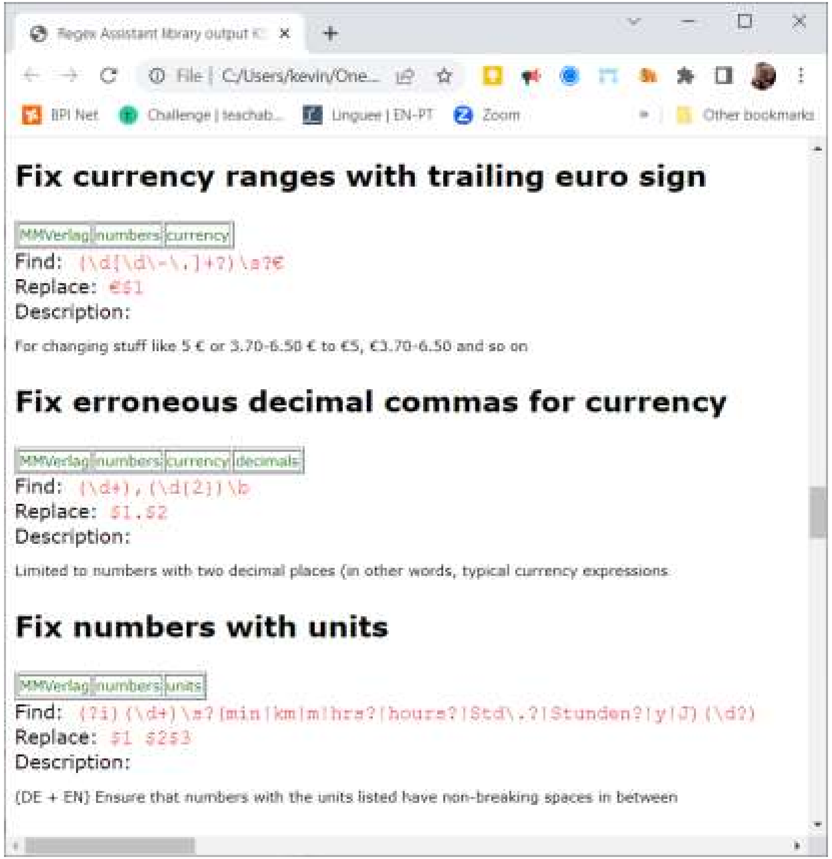
Choose File: Xmas RegexLibrary XSL.xml

File encoding: UTF-8

Option 1: Copy-paste your XSL document here (Optional if XSD referred in XML using schemaLocation)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"  xmlns:xsd="http://www.w3.org/2001/XMLSchema"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/">
<html>
```

Readable output with XSLT (3 of 3)



Fix currency ranges with trailing euro sign

MMVerlag numbers currency

Find: `(\d[\d\-\.]+?)\s?€`
Replace: `€$1`
Description:

For changing stuff like 5 € or 3.70-6.50 € to €5, €3.70-6.50 and so on

Fix erroneous decimal commas for currency

MMVerlag numbers currency decimals

Find: `(\d+),(\d{2})\b`
Replace: `$1.$2`
Description:

Limited to numbers with two decimal places (in other words, typical currency expressions

Fix numbers with units

MMVerlag numbers units

Find: `(?i)(\d+)\s?(min|km|m|hrs?|hours?|Std\.?|Stunden?|y|J)(\d?)`
Replace: `$1 $2$3`
Description:

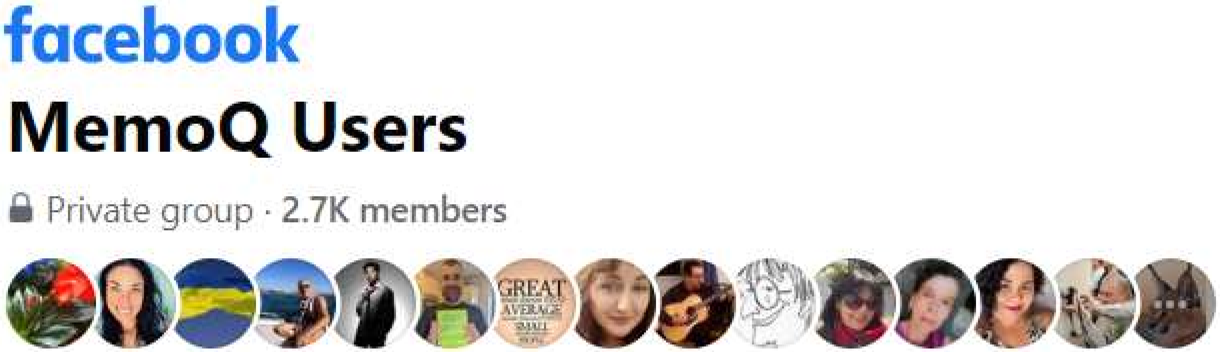
(DE + EN) Ensure that numbers with the units listed have non-breaking spaces in between

This readable list of expressions from your memoQ Regex Assistant library gives you a better overview of what you have.

Expressions in this format can be copied and pasted to use in other CAT tools such as Trados Studio, the Phrase desktop editor, etc.

XSLT can be used to create import files of regex exported from the Regex Assistant for some tools, such as TransTools+

People & places to help you



facebook

MemoQ Users

Private group · 2.7K members

When in doubt… Google it!

Online course with free resources:

memoQuickies Resource Camp (online course):
https://transtrib-tech.teachable.com/p/memoq-resource-camp

- Free section for this WTF ‘23 talk (until December 18th):
  https://transtrib-tech.teachable.com/courses/memoq-resource-camp/lectures/50698696

- Special discounted enrollment link (until December 31st):
  https://transtrib-tech.teachable.com/p/memoq-resource-camp?coupon_code=WTF23&product_id=5006468

TRANSLATION TRIBULATIONS

http://translationtribulations.com

Technology philosophy in *your* service

https://www.youtube.com/@KevinLossner

@GermanENtrans@c.im

https://www.linkedin.com/in/kslossner/

ThanQ for your attention!