

To escape, or not to escape: that is the question

Francesco Torello
Arkadia Translations S.r.l.



Arkadia in a nutshell

- Arkadia Translations is a **boutique LSP**, established in 1999
- Headquarters in **Milan** and a branch office in **Brussels**
- Core business: legal and financial translations;
- **Changed our main CAT tool in 2018**
- First time at memoQfest in 2019, I was among the speakers last year with a presentation on regular expressions and light resources



One project, many questions



- Translation of the entire website of a museum – 4 locations in different regions of Italy + series of information sheets on the most prominent artworks and museum venues



- Request for quote in mid-June 2022 – reply by the first week of July



- Website built with Adobe Experience Manager



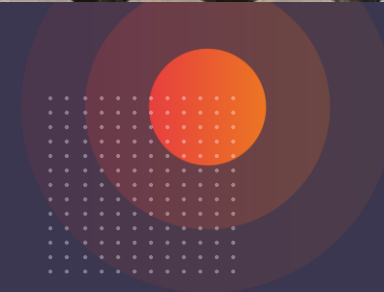
- Sample files to be analysed as soon as possible to detect any issues



- 95K words in various formats, mainly XML and Word



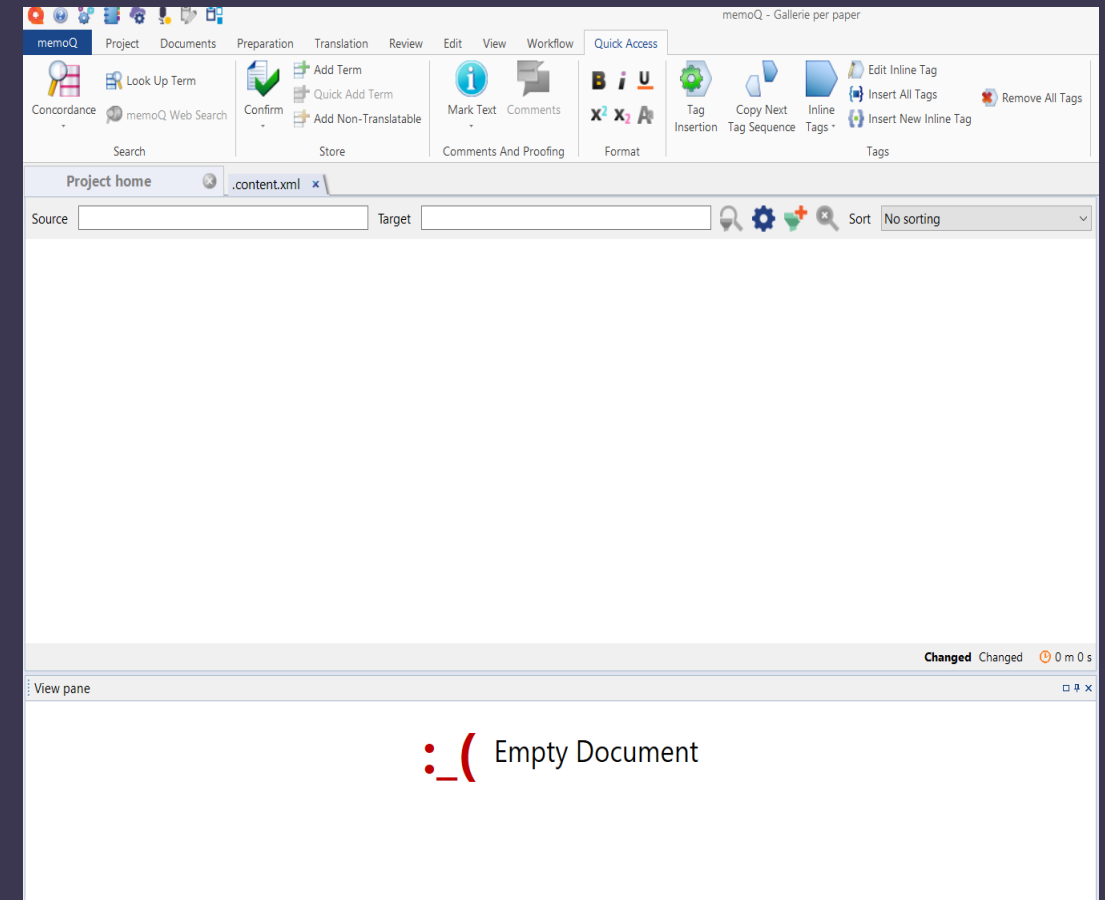
- Translation and revision between November 2022 and January 2023, with several milestones



The sample files: from the frying pan to the embers

XML sample **files**, so xml filter BUT...

1. Once imported, they turned out to be **empty**
2. **All** the sample files were **named .content.xml**, so we had to import the whole folder not to compromise the structure as our client wanted to make sure we could handle the import/export of the entire website structure
3. We had **no time** to explore the possible integration between memoQ and AEM, so we had to come up with a custom solution
4. After opening the xml files with Notepad++ we noticed that some **HTML entities** had been **escaped**



Example

```
<?xml version="1.0" encoding="UTF-8"?>
<jcr:root xmlns:sling="http://sling.apache.org/jcr/sling/1.0" xmlns:cq=
"http://www.day.com/jcr/cq/1.0" xmlns:jcr="http://www.jcp.org/jcr/1.0" xmlns:mix=
"http://www.jcp.org/jcr/mix/1.0" xmlns:nt="http://www.jcp.org/jcr/nt/1.0"
  jcr:primaryType="cq:Page">
  <jcr:content
    cq:lastModified="{Date}2022-07-19T15:32:07.162+02:00"
    cq:lastModifiedBy="U018944"
    cq:lastReplicated="{Date}2022-06-30T14:59:27.744+02:00"
    cq:lastReplicatedBy="U446583"
    cq:lastReplicationAction="Activate"
    cq:template="/conf/gdi/settings/wcm/templates/page-content"
    jcr:description="Acquista i tuoi biglietti on line e risparmia tempo. Scegli il
museo e prenota la tua visita."
    jcr:isCheckedOut="{Boolean}true"
    jcr:mixinTypes="[mix:versionable]"
    jcr:primaryType="cq:PageContent"
    jcr:title="Acquista biglietti"
    jcr:uuid="aa66142c-c7e9-4c5a-b812-d58ea5c62d06"
    sling:resourceType="gdi/components/structure/pages/page"
    brandSlug="Gallerie d'Italia"
    brandSlug_override="true"
    lastPreviewReplicate="2022-06-30T14:59:58.743+0200"
    lastPreviewReplicatedBy="U446583"
    lastPreviewReplicationAction="ACTIVATE"
    lastPreviewReplicationFailed="{Long}0"
    textIsRich="[true,true]">
```

```
<texttwocolumns
  jcr:created="{Date}2022-03-01T16:53:46.827+01:00"
  jcr:createdBy="U018464"
  jcr:lastModified="{Date}2022-04-27T12:45:42.586+02:00"
  jcr:lastModifiedBy="U018944"
  jcr:primaryType="nt:unstructured"
  sling:resourceType="gdi/components/content/textTwoColumns"
  descriptionLeft="&lt;p>Non vedi l'ora di essere a tu per tu con
l'arte?&nbsp; Acquista i tuoi biglietti on line e risparmia
tempo. Consigliata la prenotazione on
line.&nbsp;&nbsp;&lt;/p>&#xd;&#xa;&lt;p>&nbsp;&nbsp;&lt;/p>&#xd;&#xa;"
  descriptionRight="&lt;p>Dal 3 aprile 2022, ogni prima domenica del
mese ingresso gratuito per tutti i visitatori.&lt;/p>&#xd;&#xa;"
  textIsRich="[true,true]"
  title="Acquista biglietti">
  <leftCTA
    jcr:primaryType="nt:unstructured"
    ctaLayout="button_style"
    lnTarget="_self"
    showCTA="false"/>
  <rightCTA
    jcr:primaryType="nt:unstructured"
    ctaLayout="button_style"
    lnTarget="_self"
    showCTA="false"/>
</texttwocolumns>
```

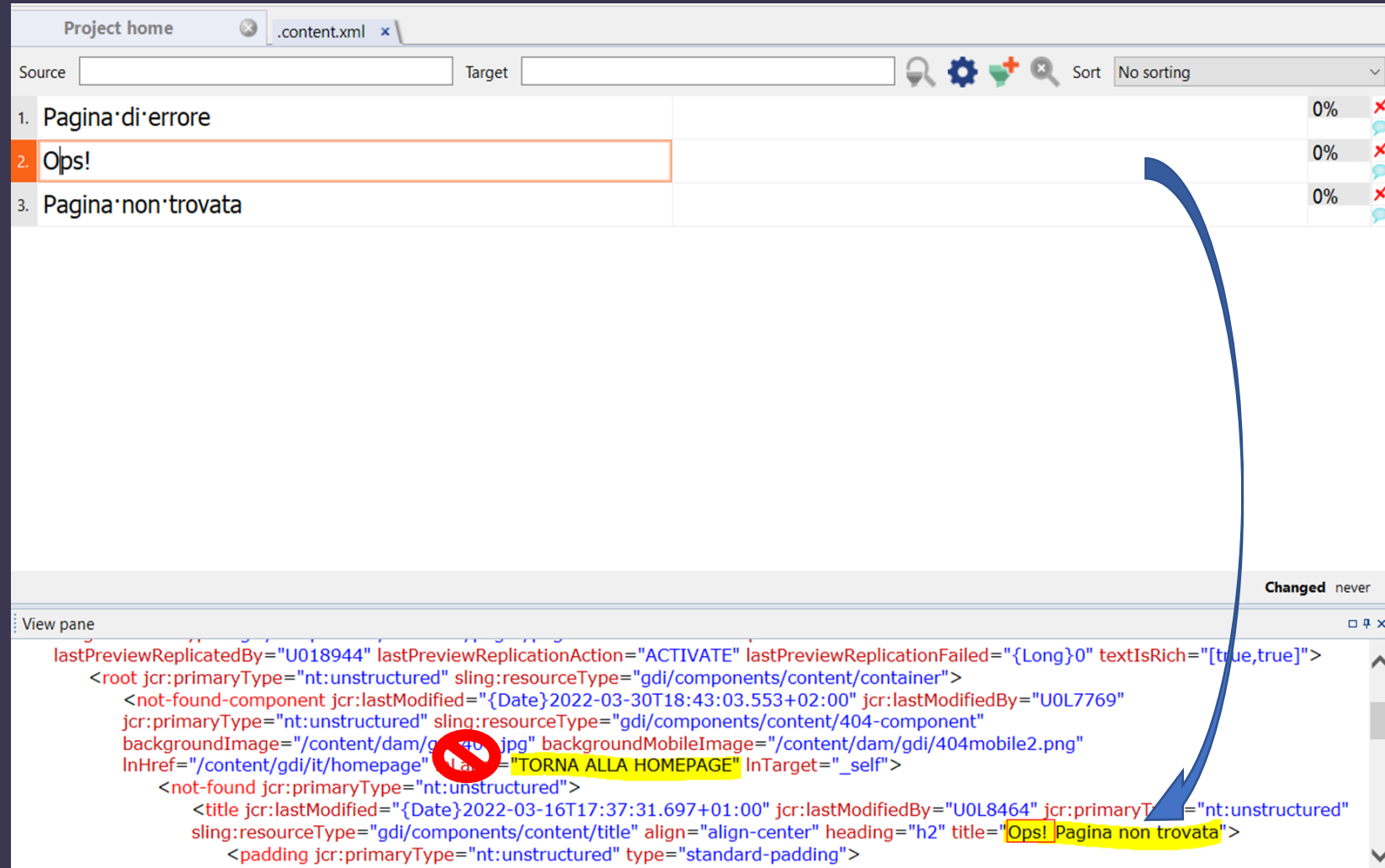



“Great things never come from comfort zones”

We had to find a solution, and find it quickly because we couldn't risk losing this prestigious customer ...

So, how can we tell memoQ what to import?

- During last year's masterclasses, we learnt about the wide range of settings and options available through **memoQ's filter configurations**, so we started playing with the most obvious one... the XML filter
- We managed to import some content, but not all of the entities were correctly handled



The screenshot shows the memoQ interface with a list of imported content items and a view pane displaying the XML representation of the selected item.

Content List:

Source	Target	Progress	Status
1. Pagina di errore		0%	✗
2. Ops!		0%	✗
3. Pagina non trovata		0%	✗

XML Representation (View pane):

```

lastPreviewReplicatedBy="U018944" lastPreviewReplicationAction="ACTIVATE" lastPreviewReplicationFailed="{Long}0" textIsRich="[true,true]">
<root jcr:primaryType="nt:unstructured" sling:resourceType="gdi/components/content/container">
  <not-found-component jcr:lastModified="{Date}2022-03-30T18:43:03.553+02:00" jcr:lastModifiedBy="U0L7769"
    jcr:primaryType="nt:unstructured" sling:resourceType="gdi/components/content/404-component"
    backgroundImage="/content/dam/gdi/404.jpg" backgroundMobileImage="/content/dam/gdi/404mobile2.png"
    InHref="/content/gdi/it/homepage" InTarget="_self">
    <not-found jcr:primaryType="nt:unstructured">
      <title jcr:lastModified="{Date}2022-03-16T17:37:31.697+01:00" jcr:lastModifiedBy="U0L8464" jcr:primaryType="nt:unstructured"
        sling:resourceType="gdi/components/content/title" align="align-center" heading="h2" title="Ops! Pagina non trovata">
        <padding jcr:primaryType="nt:unstructured" type="standard-padding">

```

A blue arrow points from the 'Ops!' item in the list to the XML representation in the view pane. A red circle with a diagonal line through it is placed over the 'Ops!' text in the XML, indicating a parsing error or a specific filter configuration.

Regex text filter – Part 1

We started exploring new options and given that

- we knew regular expressions were a possible solution
- we could use them to tell memoQ what content to import and what to exclude

We decided to use the **regex text filter**, which we had never used before...and we had no idea how to use it



Regex text filter – Part 2

We started analysing the sample files to select all the entities that contained text to be translated and came out with list of 57 entities



We then applied the simplest possible regular expression

`. * ?`

and added the other elements of the rule as in the examples shown here

`text = ". * ?"`

`title = ". * ?"`

`description = ". * ?"`

`sectionTitle = ". * ?"`

General
Paragraph
Include/exclude
Preview

☐ Rules define content to be excluded (external tags)
☒ Rules define imported content (nothing else is imported)

#	Before	Rule	After
1	text="	. * ?	"
2	title="	. * ?	"
3	description="	. * ?	"
4	ctaLabel="	. * ?	"
5	arrowLabel="	. * ?	"
6	sectionTitle="	. * ?	"
7	cities="	. * ?	"
8	city="	. * ?	"
9	descriptionLeft="	. * ?	"
10	descriptionRight="	. * ?	"
11	contactDescription="	. * ?	"

Before

text=" Rx

After

" Rx

Rule

. * ? Rx

Add
Delete
Up •
Change
Down •

So...job done? Not quite

We thought our filter worked fine and that we could start importing our sample files, as we could see sentences like this one in our editor:

```
jcr:description="Acquista i tuoi biglietti on line e risparmia tempo. Scegli il museo e prenota la tua visita."
```

BUT... we soon found out that we had also imported the escaped HTML entities and that we needed a solution to exclude them from the translatable content:

```

5. <p>Non vedi l'ora di essere a tu per tu con
   l'arte?&nbsp;Acquista i tuoi biglietti on line e
   risparmia tempo.
6. Consigliata la prenotazione on
   line.&nbsp;</p>&#xd;&#xa;<p>&nbsp;&#xd;&#xa;
7. <p>Dal 3 aprile 2022, ogni prima domenica del mese
   ingresso gratuito per tutti i visitatori.</p>&#xd;&#xa;

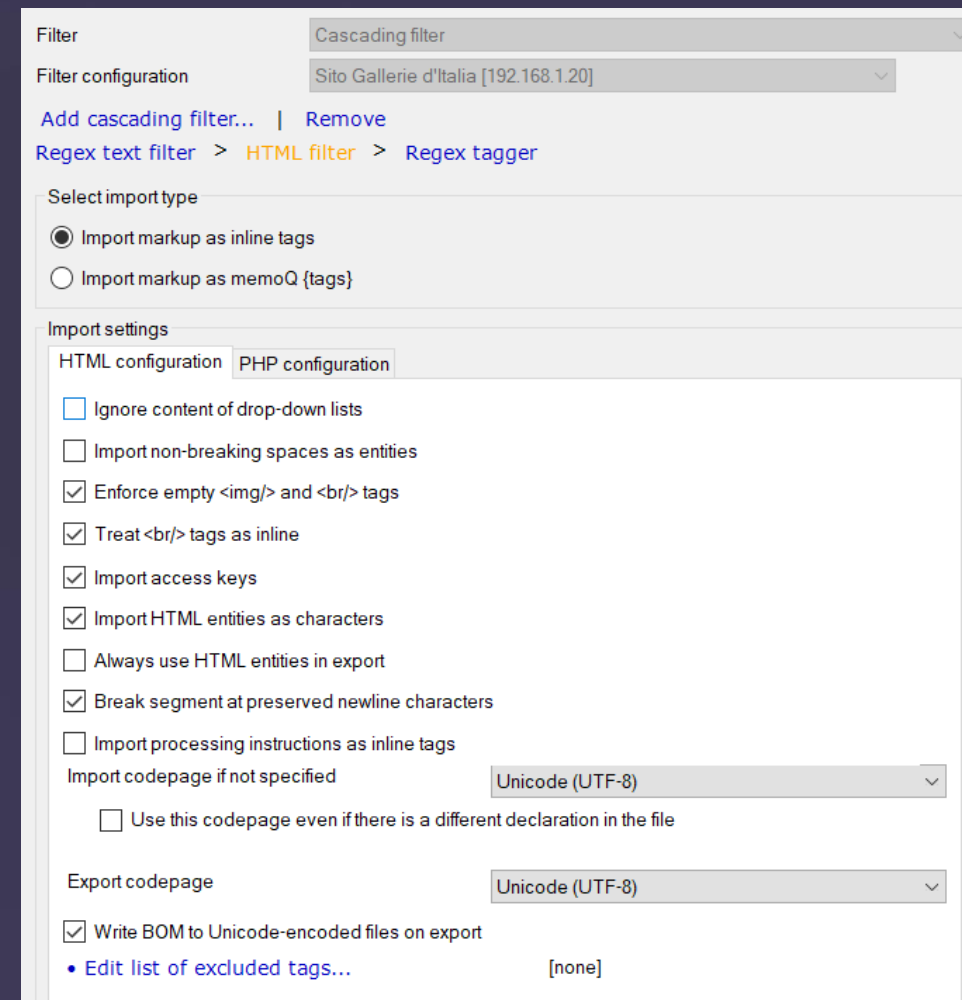
```

One filter to import them all... the cascading filter

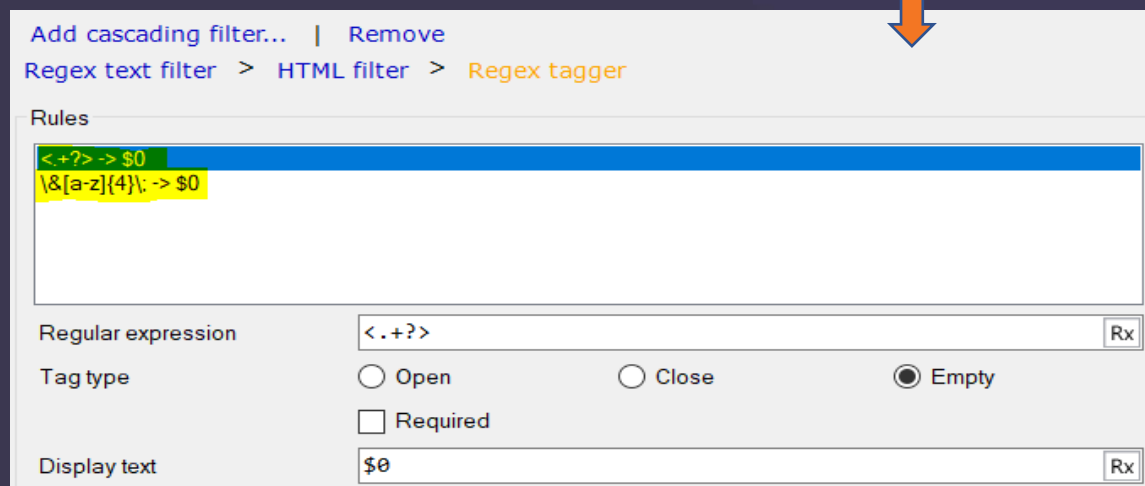
Is there a way to combine multiple filters and apply them in sequence so as to tell memoQ precisely what to import and at the same time deal with the HTML entities?

YES, the **CASCADING FILTER**

So, we developed a new cascading filter combining the regex text filter, the HTML filter and the regex tagger filter to deal with some entities that we wanted to turn into tags



The screenshot shows the 'Cascading filter' configuration window. At the top, the 'Filter' dropdown is set to 'Cascading filter' and the 'Filter configuration' dropdown is set to 'Sito Gallerie d'Italia [192.168.1.20]'. Below these, there are links for 'Add cascading filter...' and 'Remove'. A breadcrumb trail shows 'Regex text filter > HTML filter > Regex tagger'. The 'Select import type' section has two radio buttons: 'Import markup as inline tags' (selected) and 'Import markup as memoQ {tags}'. The 'Import settings' section has two tabs: 'HTML configuration' and 'PHP configuration'. Under 'HTML configuration', there are several checkboxes: 'Ignore content of drop-down lists' (unchecked), 'Import non-breaking spaces as entities' (unchecked), 'Enforce empty and
 tags' (checked), 'Treat
 tags as inline' (checked), 'Import access keys' (checked), 'Import HTML entities as characters' (checked), 'Always use HTML entities in export' (unchecked), 'Break segment at preserved newline characters' (checked), and 'Import processing instructions as inline tags' (unchecked). There are also dropdowns for 'Import codepage if not specified' and 'Export codepage', both set to 'Unicode (UTF-8)'. A checkbox 'Use this codepage even if there is a different declaration in the file' is unchecked. At the bottom, there is a checkbox 'Write BOM to Unicode-encoded files on export' (checked) and a link 'Edit list of excluded tags...' with a '[none]' value.



The screenshot shows the 'Rules' section of the 'Cascading filter' configuration window. It contains a list of rules. The first rule is highlighted in blue and has the regular expression '<.+?>->\$0'. The second rule is highlighted in yellow and has the regular expression '\\&[a-z]{4}\\:->\$0'. Below the list, there are input fields for 'Regular expression' (containing '<.+?>'), 'Tag type' (with radio buttons for 'Open', 'Close', and 'Empty', where 'Empty' is selected), 'Required' (unchecked), and 'Display text' (containing '\$0').

The escaping question

We checked the file and found out that the problem was associated with an incorrectly escaped entity.

2. "Non·si·farà·mai·più·tal·viaggio"	"No·such·voyage·will·ever·be·repeated"
3. 6·settembre·2022·-·8·gennaio·2023	6·September·2022·-·8·January·2023

In this case, " had not been escaped to

- "

```
jcr:title="Antonio Pigafetta and the first voyage around the world. "No such voyage will ever be repeated"
```

So, we

- 1) corrected the error and sent the file to the client, which gave the green light
- 2) went back to the HTML filter options to see if we could adjust something while waiting for the project to begin

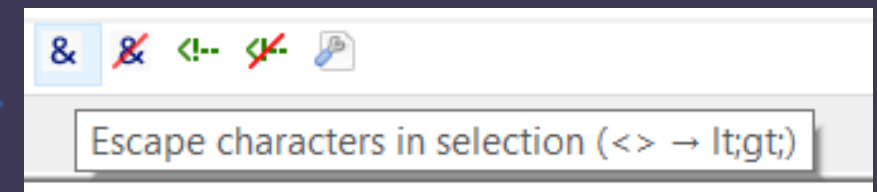
...to escape or not to escape?

We thought that maybe it was better to look for entities like &, non-breaking spaces or " in the source files and escape them before starting the project, BUT

- too many files to check (218)
- risk of deleting something due to our lack of experience with XMLs
- too time-consuming (first milestone approaching > end of November)

Eventually, we

- decided **not to escape** the entities before
- check XML syntax with Notepad++
- correct errors with the "escape" button





Conclusions

- MemoQ filters are a **very powerful** and **versatile tool**, especially when combined together in a cascading filter set
- If you **invest a reasonable amount of time** to learn how to use them, you will **save a lot of time** (and probably also **money**) later on
- **XML** or other **less user-friendly** formats may seem too much of a challenge, especially if you're not a computer geek, and **that's where memoQ filters really make the difference**. All you have to do is to **get out of your comfort zone** and learn how to use them.



Thank you! | Köszönöm!

Any questions?

f.torello@arkadiatranslations.com