



# **XML und XLIFF – wenn es nicht so benutzt wird, wie vorgesehen**

tekom/tcworld 2019

Angelika Zerfaß

[zerfass@zaac.de](mailto:zerfass@zaac.de)

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback unter

**<http://lt12.honestly.de>**

oder scannen Sie den QR-Code



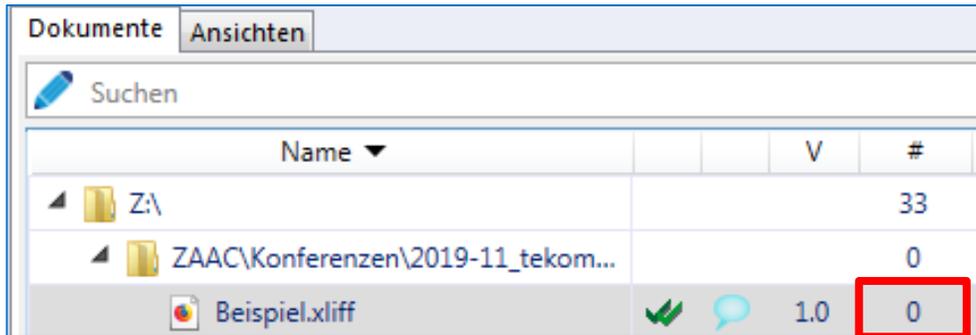
# XML und XLIFF

- Immer häufiger werden XML-Dateien oder sogar XLIFF (XML Localization Interchange File Format) in die Übersetzung gegeben.
- Mit Standardformaten sollte sich eigentlich ein einfacherer Übersetzungsprozess ergeben.
- Ganz abgesehen von den Anpassungsmöglichkeiten, die z.B. die XLIFF 1.2-Spezifikation einräumt, gibt es weitere "kreative" Wege, mit diesen Formaten umzugehen.

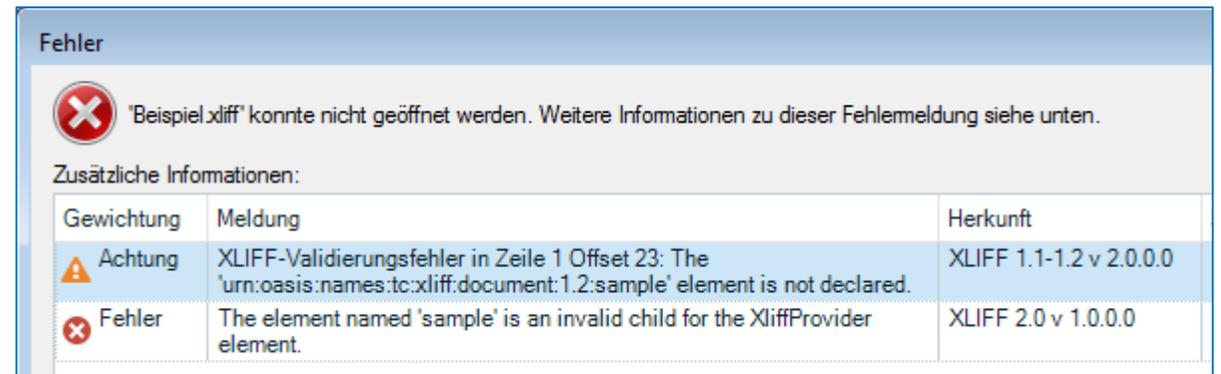
# **INKONSISTENZ: INHALT UND DATEIENDUNG**

# XML

- XML-Datei mit Erweiterung \*.XLIFF
- Import ins Übersetzungsprogramm schlägt fehl oder zeigt keine Inhalte an



Name	V	#
Z:\		33
ZAAC\Konferenzen\2019-11_tekom...		0
Beispiel.xliff	1.0	0



Fehler

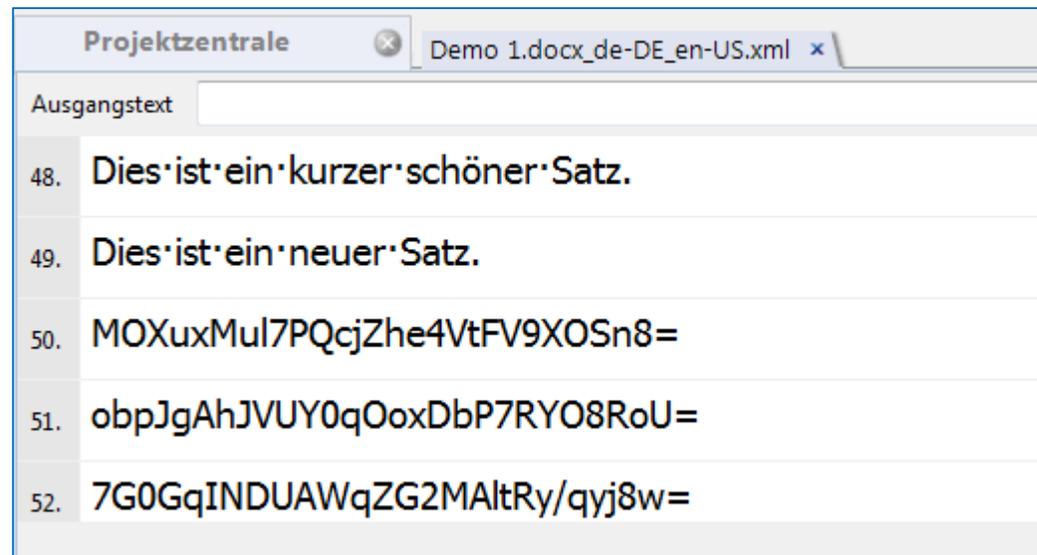
Beispiel.xliff konnte nicht geöffnet werden. Weitere Informationen zu dieser Fehlermeldung siehe unten.

Zusätzliche Informationen:

Gewichtung	Meldung	Herkunft
Achtung	XLIFF-Validierungsfehler in Zeile 1 Offset 23: The 'urn:oasis:names:tc:xliff:document:1.2:sample' element is not declared.	XLIFF 1.1-1.2 v 2.0.0.0
Fehler	The element named 'sample' is an invalid child for the XliffProvider element.	XLIFF 2.0 v 1.0.0.0

# XLIFF

- XLIFF-Datei mit Erweiterung \*.XML
  - Import ins Übersetzungsprogramm erkennt das Sprachpaar der Datei nicht und verwendet den falschen Filter.
  - Es wird mehr eingelesen, als nötig ist (bei bereits übersetzten Dateien werden sowohl Ausgangs- als auch Zieltext als Quelltext eingelesen).



# UNERWARTETE INHALTE

# XML enthält unerwartete Dinge

- Durch einen Bearbeitungsschritt, eventuell auch eine manuelle Bearbeitung, wurden Strukturen beschädigt:
  - End-Tag eines Paares ohne Anfangs-Tag
  - Überflüssige Zeichen (`<button value="Cancel" />`)
  - Nicht-übereinstimmende Tag-Namen (`<titel>How to set up a filter for XML files</title>`)
  - Tags wo keine sein dürften bzw. Abfolge von Tags nicht korrekt

# XML enthält unerwartete Dinge

- Formatierung, die man eher in einem Stylesheet erwarten würde.

```
<?xml version="1.0" encoding="Windows-1252"?>
```

```
<table_data name="contents">
```

```
<row>
```

```
<field name="ID">325610e4f320145b586245b539ef214b</field>
```

```
<field name="TITLE">Backofentage</field>
```

```
<field name="TITLE_1">Backofentage</field>
```

```
<field name="CONTENT">
```

```
[{veparse name="7cc0154b0b5fffb59896705d34c684201"}]
```

```
[row]
```

```
[col size="12" offset="0" class="col-xs-12"]
```

```
[text background color="" background image="" background fixed="" fullwidth="" class=""]
```

```
&lt;p&gt;&lt;span style="font-size: 18px;"&gt;&lt;b&gt;Vom 06.011. - 07.011. &lt;/b&gt;veranstalten wir  
Informationstage für alle Backofen-Interessenten: &lt;br&gt;&lt;br&gt;&lt;/span&gt;&lt;span  
style="font-size: 18px;"&gt;
```

```
[/text]
```

```
[/col]
```

```
[/row]
```

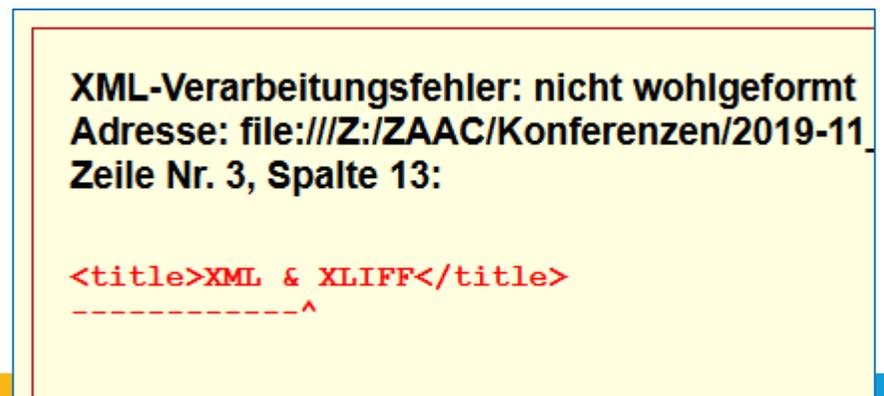
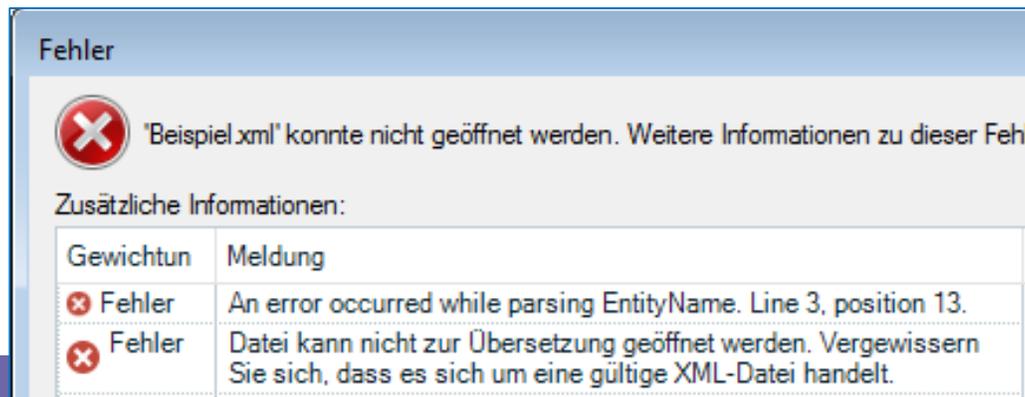
```
[{/veparse}]</field>
```

Selbsterstellte Codes in eckigen Klammern müssen über reguläre Ausdrücke getaggt werden.

HTML-Codes (&lt;p&gt;) lassen sich in TM-Systemen als Tags darstellen

# XML enthält unerwartete Dinge

- Die XML-Spezifikation definiert 5 Zeichen, die im Text als Entitäten vorkommen sollten ( " & < > ' ).
- ...For interoperability, valid documents SHOULD declare the entities amp, lt, gt, apos, quot, in the form specified in 4.6 Predefined Entities...
- "should" könnte missverstanden werden.



# XML enthält unerwartete Dinge

- Hängt vom Übersetzungsprogramm ab, ob dies als Fehler gewertet wird.
- Es gibt auch Systeme, die dies trotzdem einlesen und Zeichen im Text beim Export in die korrekte Entity überführen.

# UNGÜLTIGE ZEICHEN

# XML enthält ungültige Zeichen

✘ Fehler	": hexadecimal value 0x03, is an invalid character. Line 12892, position 31.
✘ Fehler	Datei kann nicht zur Übersetzung geöffnet werden. Vergewissern Sie sich, dass es sich um eine gültige XML-Datei handelt.

- Zum Teil stammen die Inhalte der XML-Dateien aus anderen Programmen, z.B. textbasierten Programmen, die unter anderem auch Steuerzeichen enthalten können.

```
34 <p class="subheadline">New York City, USA.</p></data>
35 <data table="tt_content" elementUid="9" key="tt_content:NEW/1/9:header"></data>
36 <data table="tt_content" elementUid="9" key="tt_content:NEW/1/9:subheader">
<![CDATA[O:49:"GridElementsTeam\Gridelements\Backend\LayoutSetup":6:{s:15:"NUL*NULrestrictions"
;N;s:14:"NUL*NULlayoutSetup";a:9:{s:15:"layoutContainer";a:10:{s:5:"title";s:16:"layout
Container";s:11:"description";s:62:"Container to get a background color or image for some}]]>
</data>
```

```
12890 <wp:meta_value><![CDATA[<h4>So erreichen Sie uns</h4>
12891 <ul class="list-unstyled list-contact">
12892 <li>Für Fragen, Wünsche und ETXAnregungen stehen wir Ihnen gerne ETXzur Verfügung. ETX</li>
12893 <li><em>Sie erreichen uns montags - freitags ETXvon 7.30 - 16.30 Uhr</em></li>
12894 </ul>]]></wp:meta_value>
```

- Manche Übersetzungsprogramme werden eine solche Datei nicht importieren.

# XML enthält ungültige Zeichen

```
<row documentation="(DE)&cr;&lf;Sichere  
Abschaltung folgender Komponenten:&cr;&lf; -  
Abschalten der 24V-Versorgung / CPV-Ventil&cr;&lf;  
- Abschalten der Digitalausgänge&cr;&lf;" .../>
```

- Auch wenn Steuerzeichen als Pseudo-Entitäten dargestellt werden, hilft das dem Übersetzungsprozess nicht viel...
- Verwendung von REGEX, um Elemente in Tags umzuwandeln.

7. (DE) Umbruch Sichere·Abschaltung·folgender·Komponenten: Umbruch ...  
Abschalten·der·24V-Versorgung·/·CPV-Ventil Umbruch ...Abschalten·der·  
Digitalausgänge...

# XML enthält ungültige Zeichen

```
<?xml version="1.0" encoding="UTF-8" ?>CRLF
<1>CRLF
  <id>6</id>CRLF
  <name>LCD-Backlight-Technologie</name>CRLF
  <nameEn>LCD-backlight-technology</nameEn>CRLF
  <typId>6</typId>CRLF
  <einheitId>2</einheitId>CRLF
  <anzeigeTyp>1</anzeigeTyp>CRLF
</1>CRLF
<2>CRLF
  <id>10</id>CRLF
  <name>Ausgabesignalsteuerung</name>CRLF
  <nameEn>output-signal-control</nameEn>CRLF
  <typId>9</typId>CRLF
  <einheitId>2</einheitId>CRLF
  <anzeigeTyp>-1</anzeigeTyp>CRLF
</2>CRLF
```

Die XML-Seite kann nicht angezeigt werden

Die XML-Eingabe kann nicht angezeigt werden, wenn Stylesheet XSL verwendet wird. Beheben Sie den Fehler und klicken Sie dann auf [Aktualisieren](#) oder wiederholen Sie den Vorgang später.

**Ein Name beginnt mit einem ungültigen Zeichen. Fehler beim Bearbeiten der Ressource**

'file:///Z:/ZAAC/Konferenzen/2019-11\_te...

```
<1>
--^
```

Elementnamen  
dürfen nicht mit  
Zahlen beginnen

# **PROBLEMATISCHE STRUKTUREN**

# XML-Strukturen nicht brauchbar

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- GXML Toolbox V3.2 R20170508; http://www.gaeb-toolbox.de/
gtcConverter.dll used version 2016.06.22 -->
<GAEB xmlns="http://www.gaeb.de/GAEB_DA_XML/DA83/3.2">
...
<BoQBody><ItemList><Item ID="IT10335800004" RNoPart="1">
...
<CompleteText><DetailTxt><Text>
<p style="font-family:Arial;font-size:10pt;text-align:left;margin-top:0pt;margin-bottom:0pt;"><span style=
"font-family:Arial;font-size:10pt;">Ger</span><span style="font-family:Arial;font-size:10pt;">ä</span><span style=
"font-family:Arial;font-size:10pt;">te, Werkzeuge und sonstige Betriebsmittel</span></p>
```

HTML-Elemente, die in XML nicht benötigt werden.

```
80. Ger ä te, Werkzeuge und sonstige Betriebsmittel
```

Vorgängerformat

```
87 ..... [StLNr]0710110711[end] CRLF
88 ..... [Langtext]{\rtf1\ansi {\fonttbl{\f0\fswiss Arial;}}{\*\generator IBBR_GXConverter
2016.06.22;}\pard\ql\f0\fs20\sb0\sa0\plain\f0\fs20\sb0\sa0 Ger\ 'e4te, Werkzeuge und
sonstige Betriebsmittel, die zur\par\pard\ql\f0\fs20\sb0\sa0\plain\f0\fs20\sb0\sa0
vertragsgem\ 'e4\ 'dfen Ausf\ 'fchrung der
```

# XML-Strukturen (problematisch)

```
<sample>
<title>This is a test sentence.</title>
<text>This sentence contains <internal>several</internal> inline <internal>tags</internal>.</text>

<text>The inline tags appear sometimes in segments that need full translation and sometimes in
segments where only the part between the inline tags needs to be translated.</text>

<text>The next segment contains the same inline tags as above, but here only the content between
the inline tags needs to be translated.</text>

<other>123456<internal>Text to translate</internal>23456<internal>Text to translate</internal>34566
</other>
</sample>
```

- Inline-Element mit zwei verschiedenen Verwendungsweisen

# Lösung

- Verwendung eines Textfilters mit regulären Ausdrücken statt eines XML-Filters
- Definition der Inhalte (Reihenfolge ist wichtig)
  1. Zwischen `<title>` und `</title>`
  2. Zwischen `<text>` und `</text>`
  3. Zwischen `<internal>` und `</internal>`
  4. 2. Filterstufe: XML

Sample_Internal with condition.xml	
1	This sentence contains <code>&lt;internal&gt;</code> several <code>&lt;/internal(...)</code> inline <code>&lt;internal&gt;</code> tags <code>&lt;/internal(...)</code> .
2	The inline tags appear sometimes in segments that need full translation and sometimes in segments where only the part between the inline tags needs to be translated.
3	Use the Regex Textfilter with a second XML filter or Regex Tagger, instead of only the XML filter
4	The next segment contains the same inline tags as above, but here only the content between the inline tags needs to be translated.
5	Text to translate
6	Text to translate
Sample_Internal with condition.xml	

# **XML MIT HTML**

# XML mit HTML

- Es gibt verschiedene Möglichkeiten, Texte in HTML-Format in XML einzubetten.
- CDATA-Bereich verwenden (nimmt die Inhalte von der Behandlung als XML aus)

```
<?xml version="1.0"?>
- <doc>
  - <para>
    <![CDATA[ <p>Translate this text.</p> ]]>
  </para>
  - <para>
    <![CDATA[ <p>Text with entities (special characters,
      like &auml; or &ouml;)</p> ]]>
  </para>
  - <para>
    <![CDATA[ <p>Text with <b>formatting</b></p> ]]>
  </para>
</doc>
```

# XML mit HTML

## ○ HTML-Strukturen über Entitäten darstellen

```
<?xml version="1.0"?>
<TEST>
<TEXT>
<para>Introduction to XYZ software</para>
<para>Please make sure that all necessary &lt;br&gt;&lt;b&gt;
extensions&lt;/b&gt; have been installed.</para>
</TEXT>
</TEST>
```

```
<?xml version="1.0"?>
- <TEST>
  - <TEXT>
    <para>Introduction to XYZ software</para>
    <para>Please make sure that all necessary
      <br><b>extensions</b> have been installed.</para>
  </TEXT>
</TEST>
```

# XML mit kreativem "HTML" ☹️

- HTML-ähnliche Strukturen

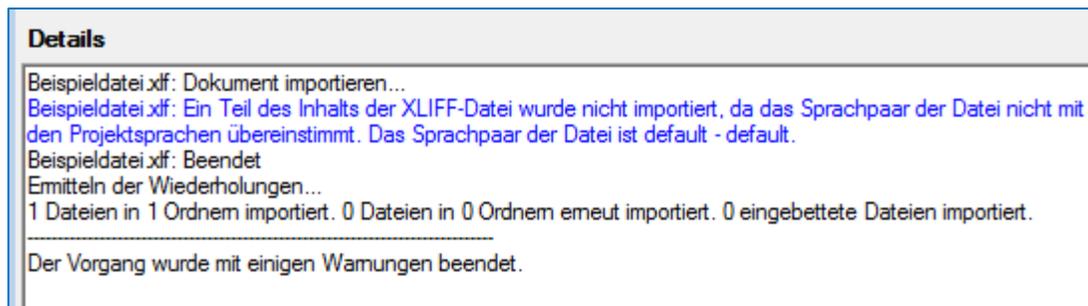
```
<?xml version="1.0"?>
- <doc>
  <para>[p]Text to translate.[/p]</para>
  <para>[p]Text with entities (special characters, like ä or ö)[/p]</para>
  <para>[p]Text with [b]formatting[/b][/p]</para>
</doc>
```

# FEHLERHAFTES XLIFF

# Fehlende Sprachangaben

- XLIFF ist zweisprachig aufgebaut und muss somit immer die Sprachangabe für Ausgangs- und Zielsprache beinhalten.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>  
<xliff version="1.2" xmlns="urn:oasis:names:tc:xliff:document:1.2">  
<file category="product" datatype="html" original="product" source-language="default" target-language="default">
```



# XLIFF-Strukturen

- Die Abfolge der Tags ist inkorrekt.

```
<?xml version="1.0" encoding="UTF-8"?>
<files>
  <madcapfile>
    <xliff>
      <file xml:space="preserve" orig...
        <header>
```

Meldung aus Programm A

Details

TYPE:  
System.Xml.XmlException

MESSAGE:  
'MadCap' ist ein nicht deklariertes Präfix. Zeile 5, Position 86.

Zusätzliche Informationen:

Gewichtung	Meldung	Herkunft
✘ Fehler	'MadCap' is an undeclared prefix. Line 5, position 86.	XLIFF 1.1-1.2 v 2.0.0.0
✘ Fehler	The element named 'files' is an invalid child for the XliffProvider element.	XLIFF 2.0 v 1.0.0.0

Meldung aus Programm B

# XLIFF mit HTML

- XLIFF, anders als XML, ist nicht gut geeignet, um HTML-codierte Inhalte zu transportieren. Die Filter der Übersetzungsprogramme erlauben es meist nicht, einen HTML-Filter nachzuschalten.
- Übersetzer muss mit HTML-Codes als Text umgehen
- HTML-Codes müssen über reguläre Ausdrücke zu Tags umgewandelt werden – was dazu führen kann, dass keine öffnenden und schließenden Tags angezeigt werden, sondern nur einzelne Platzhalter. Je nach Wissensstand desjenigen, der den Filter erstellt. (Dies erschwert dem Übersetzer die Arbeit.)
- Die Segmente sind oft auch zu groß für eine sinnvolle Bearbeitung, wenn nicht die korrekte Einstellung zur Segmentierung vorgenommen wird.



# XLIFF falsch verstanden

- Kunde schickt XLIFF-Datei mit Ausgangstext im Bereich `<source>...</source>` und wünscht, dass die Übersetzung im Bereich `<note>...</note>` eingetragen wird.
- Ein XLIFF-Filter hat keine solche Einstellung.
- Umweg über Vorbereitung (Text zwischen die `<note>`-Tags kopieren) und regulären XML-Filter.

# XLIFF mit zusätzlichen XML-Merkmalen

- Die Filter zum Einlesen von XLIFF-Dateien bieten oft nicht all die Einstellungsmöglichkeiten, die man von XML kennt.
- Beispiel:
  - XLIFF-Datei mit Angabe zur Längenbegrenzung
  - Entweder als XLIFF, dann gibt es keine Möglichkeit, die Längenbegrenzung zu importieren
  - Oder als XML, dann muss die Datei vorbereitet werden, damit nicht im Source-Bereich der Ausgangstext übersetzt wird.

```
<?xml version="1.0" encoding="utf-8"?>
<xliff version="1.2" xmlns="urn:oasis:names:tc:xliff:document:1.2">
  <file original="demo.txt" source-language="en-US" target-language="ja-jp"
  datatype="plaintext" date="2019-08-16T13:39:27Z" build-num="1.0">
    <body>
      <group id="I/O-Texts">
        <trans-unit id="TextLong_Index_10" size-unit="char" maxwidth="18">
          <source xml:space="preserve">Main Switch 1</source>
          <target xml:space="preserve">Main Switch 1</target>
        </trans-unit>
        <trans-unit id="Textshort_Index_10" size-unit="char" maxwidth="6" translate="no">
          <source xml:space="preserve">MS1</source>
          <target xml:space="preserve">MS1</target>
        </trans-unit>
        <trans-unit id="TextLong_Index_20" size-unit="char" maxwidth="20">
          <source xml:space="preserve">Master Control 2</source>
          <target xml:space="preserve">Master Control 2</target>
        </trans-unit>
        <trans-unit id="Textshort_Index_20" size-unit="char" maxwidth="6" translate="no">
          <source xml:space="preserve">MC 2</source>
          <target xml:space="preserve">MC 2</target>
        </trans-unit>
      </group>
    </body>
  </file>
</xliff>
```

# Neuestes Beispiel

```
1 <xliff version="1.0">
2 <file source-language="en" datatype="plaintext" original="messages" date=
  "2019-07-11T10:45:00Z" product-name="website">
3   <header />
4   <body>
5     <trans-unit id="translationkey" xml:space="preserve">
6       <source>Segment eins </source>
7       <target>__TRANSLATION_HERE__</target>
8     </trans-unit>
9
10    <!-- Common -->
11    <trans-unit id="common.readmore" xml:space="preserve">
12      <source>Segment zwei.</source>
13      <target>__TRANSLATION_HERE__</target>
14    </trans-unit>
15    <trans-unit id="common.email" xml:space="preserve">
16      <source>Segment drei.</source>
17      <target>__TRANSLATION_HERE__</target>
18    </trans-unit>
19
```

# Neuestes Beispiel

- Datei aus Typo3: Test-Translation.xlf.txt
  - Anweisungen vom Kunden:
    - So übersetzen Sie die HTML-Codes
    - Öffnen Sie das Dokument in einem Text-Editor
    - Übersetzen Sie zwischen den "source" Tags
    - Ersetzen Sie den Text "TRANSLATION\_HERE" mit der Übersetzung (inkonsistente Beschreibung zum Satz vorher)
    - Bitte keine Codes verändern oder löschen
  - In der Datei war die Zielsprache in dem Bereich "source-language" angegeben
  - Der Bereich "target-language" fehlte
  - Am Ende der Datei fehlten die schließenden Tags für "body", "file" und "XLIFF"
  - Im Bereiche für die Übersetzung war über all der Text "TRANSLATION\_HERE" eingetragen (was ein automatisches Einsetzen von Übersetzungsvorschlägen in Übersetzungsprogrammen verhindert).

# Zusammenfassung

- XLIFF ist XML-basiert aber nicht das gleiche, wie eine XML-Datei (zumindest aus Sicht eines Übersetzungsprogramms).
- Man sollte das Dateiformat, das man verwendet einigermaßen verstehen. 😊
- Ein Testlauf durch ein TM-Programm ist angeraten.
- Auch Übersetzungsanbieter müssen XML und XLIFF verstehen, um geeignete Einstellungen oder gegebenenfalls Vorbereitungsschritte vornehmen zu können bzw. den Kunden auf Fehler im Dateiformat hinweisen zu können.



Angelika Zerfaß  
[zerfass@zaac.de](mailto:zerfass@zaac.de)

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback unter <http://lt12.honestly.de> oder scannen Sie den QR-Code

