

STANDARD FORMATS IN TRANSLATION

TMX, TBX AND XLIFF

WHAT DATA CAN BE TRANSFERRED WITH THESE FORMATS?

Angelika Zerfaß

The Localization and Translation Conference, Warsaw 2015

Agenda

- TMX
 - Translation Memory Exchange format
- TBX
 - Term Base Exchange format
- XLIFF
 - XML Localization Interchange File format
(bilingual translation format)

TMX

- Contains the segment pairs of a translation memory.
- Contains the system data (user name, save date...) for the segment pairs.
- Can contain user-defined data that categorizes the segment pairs (project numbers, client names...)
- Can contain tool-specific data

TMX

- Being an exchange format, you would expect that all information in a TMX file can be exchanged between tools.
- But when you try it out, you will find that this mostly applies to the segment pairs themselves and the system data. Any user-defined or tool-specific data will be lost.

TMX

- A closer look at the metadata categories that can be saved to a TM:

Field Values	
Custom Fields	
Field	Value
Doc Type	manual,website
Project Number	12345
Department	ABC
Domain	education

Custom Fields	
Field	Value
Quality	100
SourceFile	Demo1_de_Demo1_en...
TargetFile	Demo1_de_Demo1_en...

System Fields	
Field	Value
Created by	TOOLS\angelikazerfass
Created on	24.02.2015 15:21:43
Document structure	x-tm-tag
Last modified by	TOOLS\angelikazerfass
Last modified on	24.02.2015 15:21:43
Last used by	TOOLS\angelikazerfass
Last used on	24.02.2015 15:23:19
Usage count	1

SDL Trados Studio 2014

Segment details	
Created by	angelikazerfass
Created at	24.02.2015 15:25:51
Modified by	angelikazerfass
Modified at	24.02.2015 15:25:51
Last modified role	None
Document	3 sentences.docx
Corrected	No
Aligned	No

Metadata	
Project	Training
Client	Client A
Domain	Education
Subject	Training Manual

Context	
Previous segment	This is the first sentence.
Next segment	This is the third sentence.
Context ID	

Custom fields	
Translator Name	
Doc Type	marketing legal website manual
Department	ABC
Project Number	12345

memoQ 2014

Example of TM data (Trados Studio / memoQ)

Field Values	
Custom Fields	
Field	Value
Doc Type	manual.website
Project Number	12345
Department	ABC
Domain	education

Segment details

Created by	angelikazerfass
Created at	24.02.2015 15:25:51
Modified by	angelikazerfass
Modified at	24.02.2015 15:25:51

System data (data that is saved by the TM tool automatically when you save a segment pair to the TM).

Exchange between tools possible without loss.

System Fields	
Field	Value
Created by	TOOLS\angelikazerfass
Created on	24.02.2015 15:21:43
Document structure	x-tm-tag
Last modified by	TOOLS\angelikazerfass
Last modified on	24.02.2015 15:21:43
Last used by	TOOLS\angelikazerfass
Last used on	24.02.2015 15:23:19
Usage count	1

Last modified role	None
Document	3 sentences.docx
Corrected	No
Aligned	No

Metadata

Training
Client A
Education
Training Manual

Context

Previous segment	This is the first sentence.
Next segment	This is the third sentence.
Context ID	

Custom fields

Translator Name	
Doc Type	marketing legal website manual
Department	ABC
Project Number	12345

Example of TM data (Trados Studio / memoQ)

Field Values	
Custom Fields	
Field	Value
Doc Type	manual,website
Project Number	12345
Department	ABC
Domain	education

User-defined fields can be created as text fields and list fields (list of predefined values and are saved with the translations to the TM).

Exchange of this kind of data is possible, but depends on the combination of tools you exchange between.

System Fields	
Field	Value
Created by	TOOLS\angelikazerfass
Created on	24.02.2015 15:21:43
Document structure	x-tm-tag
Last modified by	TOOLS\angelikazerfass
Last modified on	24.02.2015 15:21:43
Last used by	TOOLS\angelikazerfass
Last used on	24.02.2015 15:23:19
Usage count	1

Segment details	
Created by	angelikazerfass
Created at	24.02.2015 15:25:51
Modified by	angelikazerfass
Modified at	24.02.2015 15:25:51
Last modified role	None
Document	3 sentences.docx
Corrected	No
Aligned	No
Metadata	
Project	Training
Client	Client A
Domain	Education
Subject	Training Manual
Context	
Previous segment	This is the first sentence.
Next segment	This is the third sentence.
Context ID	
Custom fields	
Translator Name	
Doc Type	marketing legal website manual
Department	ABC
Project Number	12345

Example of TM data (Trados Studio / memoQ)

Field Values	
Custom Fields	
Field	Value
Doc Type	manual.website
Project Number	12345
Department	ABC
Domain	education

Tool-specific data:
 Studio: document structure (was the segment a heading, link, footnote...)
 Dates when and by who a segment was last used as it is and not changed as well as how often it was used.

System Fields	
Field	Value
Created by	TOOLS\angelikazerfass
Created on	24.02.2015 15:21:43
Document structure	x-tm-tag
Last modified by	TOOLS\angelikazerfass
Last modified on	24.02.2015 15:21:43
Last used by	TOOLS\angelikazerfass
Last used on	24.02.2015 15:23:19
Usage count	1

Segment details

Created by	angelikazerfass
Created at	24.02.2015 15:25:51
Modified by	angelikazerfass
Modified at	24.02.2015 15:25:51
Last modified role	None
Document	3 sentences.docx
Corrected	No
Aligned	No

Tool-specific data:
 Modification role (user of an online project who last saved this segment pair – translator, reviewer1, reviewer2, admin)
 Document name
 Has the source text been edited before saving to the TM?

Context

Previous segment	This is the first sentence.
Next segment	This is the third sentence.
Context ID	

Custom fields

Explicit context information (source sentence before and after the segment pair) and maybe a context ID, if a file format containing such IDs, for example software strings in Excel or XML, has been specified during import.

Department	ABC
Project Number	12345

Example of TM data (Trados Studio / memoQ)

Field Values	
Custom Fields	
Field	Value
Quality	100
SourceFile	Demo1_de_Demo1_en...
TargetFile	Demo1_de_Demo1_en...

Segment details

Created by	angelikazerfass
Created at	24.02.2015 15:25:51
Modified by	angelikazerfass
Modified at	24.02.2015 15:25:51
Last modified role	None
Document	3 sentences.docx
Corrected	No
Aligned	No

Metadata

Project	Training
Client	Client A
Domain	Education
Subject	Training Manual

Context

Previous segment	This is the first sentence.
Next segment	This is the third sentence.
Context ID	

Custom fields

Translator Name	
Doc Type	marketing legal website manual
Department	ABC
Project Number	12345

Information whether the segment pair comes from an alignment.

Studio: connection between the segments (100 = confirmed) plus the name of the source and target documents of the alignment

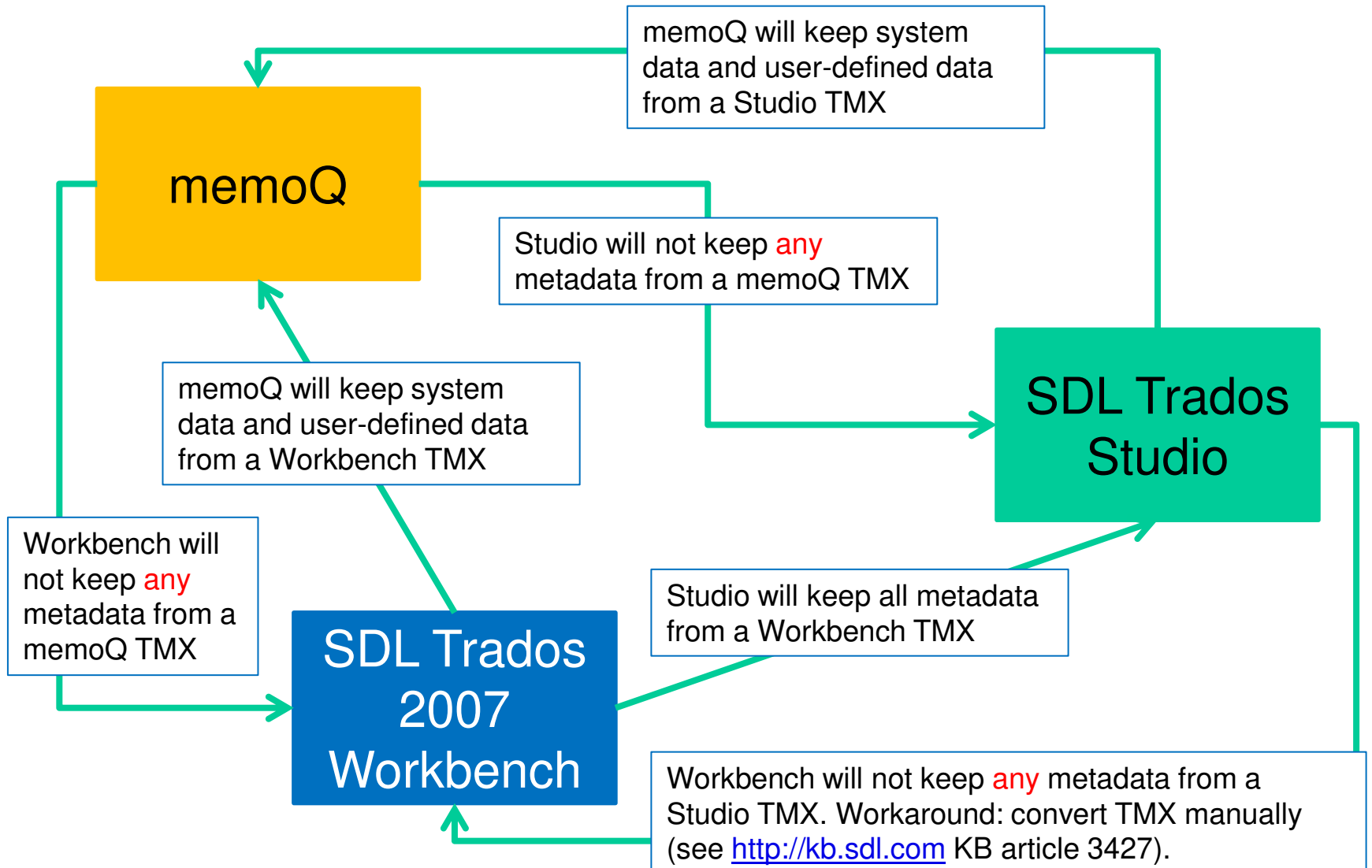
memoQ: Alignment yes/no

Information on alignment can be used to apply penalties on the match values during translation.

This information cannot be reused after TMX exchange.

Last modified on	24.02.2015 15:21:43
Last used by	TOOLS\angelikazerfass
Last used on	24.02.2015 15:23:19
Usage count	1

Metadata during import of TMX



TMX Exchange Test

- Translate HTML, DOCX and IDML with Studio and memoQ.
- Export TMX from both tools and import into other tool.
- Analyze match values from TM created with TMX.
- Test is not representative as the samples were very small (20 segments each) and contained a lot of formatting, tabs, breaks, index entries...

TMX Exchange Test

	A	B	C
1	TMX from Studio to memoQ	best results	
2			
3	HTML	57% of 100% matches	
4	DOCX	45% of 100% matches	43% high fuzzy matches (75%-99%)
5	IDML	73% of 100% matches	27% high fuzzy matches (75%-99%)
6			
7			
8			
9	TMX from memoQ to Studio	best results	
10			
11	HTML	80% of 100% matches	19% high fuzzy matches (75%-99%)
12	DOCX	56% of 100% matches	24% high fuzzy matches (75%-99%)
13	IDML	35% of 100% matches	50% high fuzzy matches (75%-99%)
14			
15			
16	SDL Trados Studio 2014 / memoQ 2014		
17			

TMX Exchange Test

- Why do we get different match values after a TMX exchange?
 - Segmentation can be different and therefore the match from a TMX exchange might not fit any longer.
 - Tools will see text differently (memoQ extracts text from attributes in HTML by default, Studio only does so if the filter is adjusted)
 - Matches will differ if the sentence contains special elements like index entries, tabs, automatic fields (Word, like CurrentDate)...
 - Penalties on alignment segments cannot be set, because the receiving tool does not know that a segment came from an alignment.

TMX Exchange Test

- Why do we get no context matches after a TMX exchange?
 - Context matches are saved in different ways in the different TM tools
 - Studio: Hash code that consists of information about the previous segment, translation, document structure (heading, link, paragraph...)
 - memoQ: explicit segment before and after the saved segment pair

How Studio saves context information to the TMX file:

```
<prop type="x-Context">2977040540754490337, -2182033961215568804</prop>
```

How memoQ saves context information to the TMX file:

```
<prop type="x-context-pre">&lt;seg&gt;previous sentence&lt;/seg&gt;</prop>  
<prop type="x-context-post">&lt;seg&gt;following sentence&lt;/seg&gt;</prop>
```

TBX

- TBX is a standard exchange format for term base content.
- Not all translation tools support TBX as import and/or export format.
 - SDL MultiTerm: TBX export / TBX import via conversion with MultiTerm Convert
 - Across: TBX import / TBX export
 - memoQ: TBX import (for the fields that are available in the term base module / no export to TBX from internal term base module, but import and export available in qTerm (web-based term base))

```

<tu tuid="1" datatype="Text" srclang="en">
  <tuv xml:lang="en-us">
    <seg>This is a test.</seg>
  </tuv>
  <tuv xml:lang="de">
    <seg>Dies ist ein Test.</seg>
  </tuv>
  <tuv xml:lang="ja">
    <seg>テストです。</seg>
  </tuv>
  <tuv xml:lang="zh-cn">
    <seg>这是试验。</seg>
  </tuv>
</tu>

```

TMX

```

- <termEntry id="c2">
  - <descrip type="subjectField">
    Hardware \ Other Processing Units and Specialized Devices
  </descrip>
  <descrip type="relatedConceptBroader">acceptor</descrip>
  - <langSet xml:lang="en">
    <admin type="productSubset">Retail Store Solutions</admin>
    - <adminGrp>
      <admin type="sourceIdentifier">Translation Services Center</admin>
    </adminGrp>
    - <ntig>
      - <termGrp>
        <term>bill acceptor</term>
        <termNote type="partOfSpeech">noun</termNote>
      </termGrp>
      - <descrip type="context">
        Accepts bill denominations of $1, $2, $5, $10, $20, $50 and $100. The bill acceptor c
        holds 600 bills. It detects and rejects counterfeit bills.
      </descrip>
    </ntig>
  </langSet>
  - <langSet xml:lang="fr">
    <admin type="productSubset">Retail Store Solutions</admin>
    - <ntig>
      - <termGrp>
        <term>accepteur de billets</term>
        <termNote type="partOfSpeech">nom</termNote>
      </termGrp>
    </ntig>
  </langSet>
</termEntry>

```

TBX

Global information in entry head

Language ID

Administrative data of this language

Term in English

Information on term level

Language ID

Term in French

```
- <termEntry id="c2">
```

```
- <descrip type="subjectField">
```

```
Hardware \ Other Processing Units and Specialized Devices
```

```
</descrip>
```

```
<descrip type="relatedConceptBroader">acceptor</descrip>
```

```
- <langSet xml:lang="en">
```

```
<admin type="productSubset">Retail Store Solutions</admin>
```

```
- <adminGrp>
```

```
<admin type="sourceIdentifier">Translation Services Center</admin>
```

```
</adminGrp>
```

```
- <ntig>
```

```
- <termGrp>
```

```
<term>bill acceptor</term>
```

```
<termNote type="partOfSpeech">noun</termNote>
```

```
</termGrp>
```

```
- <descrip type="context">
```

```
Accepts bill denominations of $1, $2, $5, $10, $20, $50 and $100. The bill acceptor holds 600 bills. It detects and rejects counterfeit bills.
```

```
</descrip>
```

```
</ntig>
```

```
</langSet>
```

```
- <langSet xml:lang="fr">
```

```
<admin type="productSubset">Retail Store Solutions</admin>
```

```
- <ntig>
```

```
- <termGrp>
```

```
<term>accepteur de billets</term>
```

```
<termNote type="partOfSpeech">nom</termNote>
```

```
</termGrp>
```

```
</ntig>
```

```
</langSet>
```

```
</termEntry>
```

TBX

- During export, a tool will create a TBX structure.
- During import the fields in the TBX file will have to be assigned to the available fields in the receiving term base system.

Obstacles to TBX exchange

- Term base components of TM tools can have very different functionalities.
- They range from fixed-layout term bases to term-bases that allow some additional user-defined fields to term bases that are freely configurable.
- This means that the term base structures are very diverse.
 - A term base system that has a fixed layout obviously will not be able to import content of fields that do not exist in the term base.

Terminology Exchange

- If both tools support TBX, this should be the preferred way of exchanging terminology.
- If not, a delimited or table-based format might be easier to handle, but will probably not be able to transport all data from one system to the other.

XLIFF

- XLIFF (XML Localization Interchange File Format) was created to
 - be the single format for translation (independent of the source format of the file)
 - be a bilingual file format that holds the source segments as well as the target segments
 - be a file format that can hold a lot of metadata and additional information about the segments, like
 - the history of a segment with its changes
 - the status of a segment (confirmed, proofread, rejected)
 - where the match came from (name of the TM or MT system)
 - comments that were set in the translation tool
 - ...

XLIFF

- Some tools have adopted XLIFF as their internal file format (MQXLIFF, SDLXLIFF...), but as the XLIFF specification allows a lot of customization, the exchange of XLIFF is not without drawbacks.
- Example:
 - An XLIFF file is prepared in tool A and sent to a user with tool B.
 - The user translates the file and sets comments.
 - The file is sent back to tool A.
 - The user of tool A cannot see the comments, because the way the comments are incorporated into the XLIFF files is different for tool A and B.

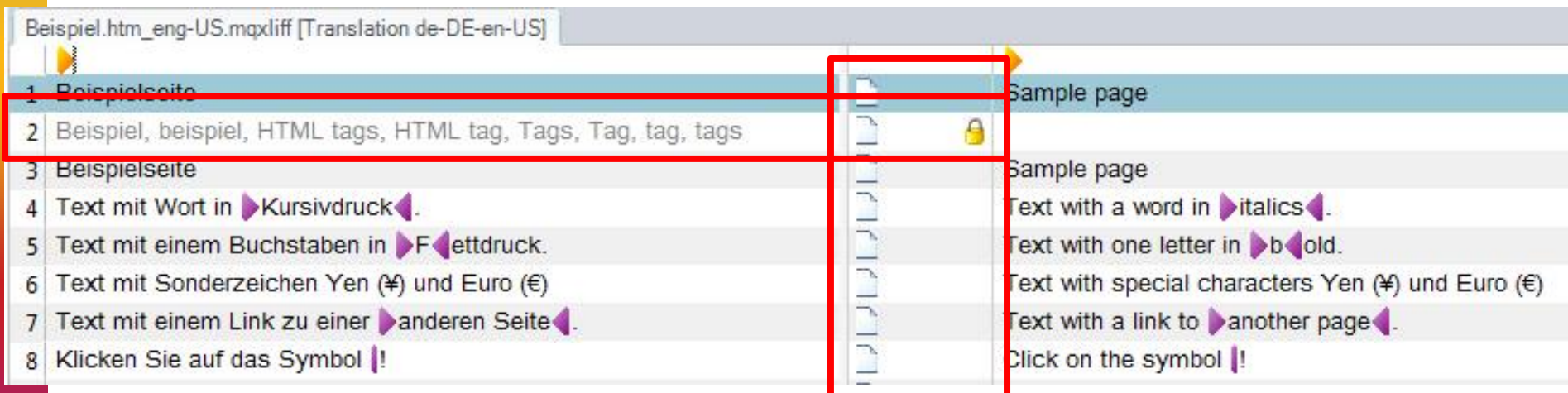
memoQ

- Segment with comment (yellow bubble)
- Locked segment (gray line / lock symbol)
- Rejected segment (status field red)
- Segment confirmed as reviewer (double checkmark)

Source	Target	Progress	Status
1. Beispielseite	Sample page	0%	Confirmed (double checkmark)
2. Beispiel, beispiel, HTML tags, HTML tag, Tags, Tag, tag, tags		0%	Locked (lock icon)
3. Beispielseite	Sample page	0%	Confirmed (double checkmark)
4. Text mit Wort in <i>Kursivdruck</i> .	Text with a word in <i>italics</i> .	0%	Comment (yellow bubble)
5. Text mit einem Buchstaben in F ett druck.	Text with one letter in b old .	100%	Confirmed (double checkmark)
6. Text mit Sonderzeichen Yen (¥) und Euro (€)	Text with special characters Yen (¥) und Euro (€)	71%	Rejected (red bar)
7. Text mit einem Link zu einer anderen Seite <a>.	Text with a link to <a>another page .	100%	Confirmed (double checkmark)
8. Klicken Sie auf das Symbol !	Click on the symbol !	100%	Confirmed (double checkmark)

When opening this XLIFF file from memoQ in Studio:

- All segments appear as not edited/confirmed
- Comment is not visible
- Locked segment appears as locked in Studio as well
- Rejected segment does not appear as rejected
- Status of segment that was confirmed as reviewer not visible
- Match values not visible



SDL Trados Studio

- Segment with comment (highlighted text)
- Locked segment (segment grayed out / lock symbol)
- Rejected segment (reject symbol)
- Segment confirmed by reviewer (reviewer confirmed symbol)

Beispiel.htm_de-DE_en-US.sdlxliff [Review]*		
1	Beispielseite	Sample page
2	Beispielseite	Sample page
3	Text mit Wort in <i>Kursivdruck</i> .	Text with a word in <i>italics</i> .
4	Text mit einem Buchstaben in F ettdruck.	Text with a letter in b old.
5	Text mit Sonderzeichen Yen (¥) und Euro (€).	Text with special characters Yen (¥) and Euro (€).
6	Text mit einem Link zu einer anderen Seite .	Text with a link to another page .
7	Klicken Sie auf das Symbol !	Click on the ! symbol!
8	Kreise rot-schwarz	Red-black circles.

When opening this XLIFF file from Studio in memoQ:

- Comment is visible (additional comments for segments that have a different confirmation status)
- Locked segment appears as locked in Studio as well
- Rejected segment status visible
- Reviewed segment status visible
- Match values visible

Source	Target	Progress	Status
1. Beispielseite	Sample page	0%	Confirmed
2. Beispielseite	Sample page	100%	Locked
3. Text mit Wort in g Kursivdruck g .	Text with a word in g italics g .	0%	Confirmed
4. Text mit einem Buchstaben in g F g ettdruck.	Text with a letter in g b g old.	87%	Confirmed
5. Text mit Sonderzeichen Yen (x) und Euro (€)	Text with special characters Yen (x) and Euro (€)	0%	Rejected
6. Text mit einem Link zu einer g anderen Seite g .	Text with a link to g another page g .	0%	Confirmed
7. Klicken Sie auf das Symbol x !	Click on the x symbol!	0%	Confirmed
8. Kreise rot-schwarz	Red-black circles.	0%	Rejected

Comments will not be visible any more, when the file goes back to Studio, but locking information, rejection/confirmation status will still be there.

XLIFF

- Other tools might also create XLIFF files for translation.
- Unfortunately they are not always usable with the tools because their setup is incomplete.
- This is how an XLIFF file could look like:

```
<?xml version="1.0"?>
<xliff version="1.1">
```

```
...
```

```
<body>
```

```
<trans-unit id="1">
```

```
<source xml:lang="EN-US">XLIFF Tool</source>
```

```
<target xml:lang="DE-DE">XLIFF Tool</target>
```

```
</trans-unit>
```

```
<trans-unit id="3">
```

```
<source xml:lang="EN-US">XLIFF processing</source>
```

```
<target xml:lang="DE-DE"></target>
```

```
</trans-unit>
```

```
<trans-unit id="4">
```

```
<source xml:lang="EN-US">XLIFF Data Manag
```

```
<target xml:lang="DE-DE">XLIFF Datenmanag
```

```
</trans-unit>
```

```
</body>
```

```
</xliff>
```

Copy of the source text between the target tags.

Area between target tags is empty. The translation of the source text will appear here after processing.

The target area already contains a translation.

XLIFF


- Ideally, any text that already exists as translation is marked with additional information inside the tag, like `TRANSLATED`, which can be used by translation tools to exclude already translated segments.
- An XLIFF file that does not contain the target tags cannot be used by translation tools.
- An XLIFF file where the translatable text is anywhere else than between the source tags cannot be used by translation tools.

XLIFF

- XLIFF 2.0 is the latest version of this standard file format and is supposed to make the exchange of additional data easier as it does not allow as many customizations for crucial information.

Summary

- Standard formats allow exchange between different systems, but usually only up to a certain point.
- Different tools have different ways to save their data and as most standard formats allow user-defined extensions, a complete exchange of all data and metadata is not possible.



Thank you for
your attention

zerfass@zaac.de