Best Practice for Leveraging Legacy Translation Memory when Migrating to DITA

Gershon Joseph

Tech-Tav Documentation Ltd.

Rodolfo Raya

Heartsome Holdings Pte. Ltd.

1. Statement of Problem

Many organizations have previously translated content that was authored in non-DITA tools (such as Word and FrameMaker). When migrating their legacy content into the new DITA authoring environment, what does the organization do about their legacy translation memory? This legacy translation memory (TM) was created with large financial investment that can't easily be thrown away simply because a new authoring architecture is being adopted.

This article describes best practices that will help organizations to use their legacy TM for future translation projects that are authored in DITA, in order to minimize the expense of translating DITA-based content.

2. Terminology

Before we get into the details, let's define the terms used in the localization industry so that subsequent sections will be better understood.

CAT

> **Computer Aided Translation**, which helps the translator translate the source content. CAT tools usually leverage Translation Memory to match sentences and inline phrases that were previously translated. In addition, some CAT tools use Machine Translation to translate glossary and other company-specific terms (extracted from a terminology database).

CMS

> **Content Management System**, while help teams manage, store, version and publish their source content.

Matching

> The level of accuracy with which CAT tools can match content being translated to the TM. The levels of matching are defined as follows:

Fuzzy matching

> The source segment being matched is similar, but not identical to, the source language segment in the TM.

Leveraged matching

> The source segment being matched is identical to the matched segment, but the context is not known.

Exact matching

> The source segment being matched is identical to the matched segment and comes from exactly the same context.

MT

> **Machine Translation** is a technology that translates content directly from source without human intervention. Used in isolation, MT usually generates an unusable translation. However, when integrated into a CAT tool to translate specific terminology, MT is a useful technology.

TM

> **Translation Memory** is a technology that reuses translations previously stored in the database used by the translation tool. TM preserves the translation output for reuse with subsequent translations.

TMX

> **Translation Memory eXchange** is an industry standard format for exchanging TM between

CAT tools.

XLIFF

XML Localisation Interchange File Format is a document format used to exchange translatable content between CAT tools.

3. Recommended Best Practices

3.1. General Process

This section describes the process at a high level that is independent of tools used and the features they support.

If DITA content is stored on a file system or in a CMS that does not support TMX (or in a CMS that uses proprietary techniques to implement reuse and other features that would be lost by exporting to TMX), then proceed as follows:

1. Migrate the legacy content to DITA.
2. Modify the DITA source as necessary for the document release to be translated.
3. When the DITA source has been approved in the source language, do one of the following:
    - Transform the DITA content into XLIFF, which can handle segments at the block or sentence level. This is the recommended option if supported by your DITA tool chain.
    - If your tool chain does not support export to XLIFF, send a copy of the DITA source to your translation service provider.

    ### Note

    If your CMS provides the translator direct access to the DITA XML, then the translator uses the standard techniques provided by the CMS to access and translate the DITA content.

4. The translator uses standard methodology for the translation tool to translate the DITA content, leveraging the legacy TM. The following points should be kept in mind when translating DITA content:
    - Provided the structure of the DITA-based content has not changed radically compared to the legacy documents, the CAT software should achieve exact matching on most segments in the TM. As long as the legacy TM aligns with the DITA source at the sentence level, the translation software should be able to achieve leveraged matching for the elements. Most CAT tools break the DITA block elements down into sentence-level segments, which will ensure better matching of the legacy TM.

  - Inline elements may not match at all, or may only be fuzzy matches. If a CAT tool is used to preprocess the TM to prepare it for the DITA-based translation project, then inline elements should yield an exact match. Note that the TM engine should help you recover 70% of the inline tags, which is the main area where matching is prone to fail.
  - If conrefs are used as containers for reusable text, then these items may not exactly match (only fuzzy match at best). However, since each of these items needs to be translated only once, and should at least fuzzy match, it should not result in significant translation expense. For best practices on using conref elements in DITA documents that need to be translated, please see XREF TO CONREF BEST PRACTICE.
  - When text entities are used as containers for reusable text, it is preferable to use a CAT tool that extracts translatable text from the XML files using an XML parser. The XML parser will insert the content of the text entities into the source text that the translator uses as a reference. This allows the translator to check that the translated segments flow correctly in the target language. If text entities are translated separately from the context where they are used, there may be grammatical inconsistencies in the final text when the translated DITA files are published.

5. After the translated content has been approved, it can be published using the same publishing tools used to publish the source language DITA content.

3.2. Using DITA with TMX

This section describes the recommend approach when using tools that fully support DITA and TMX.

When your DITA content is ready to be translated for the first time, do the following:

1. Export the legacy TM to TMX and tweak the segmentation rules. This step is optional.

   This process of creating a better aligned TM should result in an improvement of 10-20% on TM matching. Whether it's worth the effort and expense in doing this process depends on the size of the DITA documents to be translated and the number of target languages. If the number of target languages is small, it may be more economical to retranslate fuzzy matches in a separate file. However, if the word count is high and there are many target languages, tuning the TM will always yield substantial translation savings.

   Proceed as follows:

   a. Export the legacy TM to a TMX file.
   b. Tune the segmentation rules.

      The TMX file is an XML file, which can be manipulated to better align the translation segments with the DITA markup.

When tuning your legacy TM, take the following into account:

- *Unmatched tags* — Unmatched tags can result from conditional text marked up in legacy tools (such as FrameMaker), or when block elements contain several sentences that share a common format marker (for example, a paragraph containing several sentences marked as bold; the first sentence contains only an opening bold tag, and the last sentence contains only a closing bold tag).
- *Segmentation rules* — The segmentation rules used for translating legacy material may not be well suited for XML documents. For example, your legacy Word or FrameMaker-based segmentation rules may include a rule to terminate a segment after a colon, to separate a procedure title from the steps. Since DITA uses markup to indicate where the procedure title ends and the steps begin, this segmentation rule can be discarded.

2. Convert the modified TMX file back into a TM.

   This new TM will provide more exact matching against your DITA content than the legacy TM.

- Export the DITA documents to XLIFF.
- Import the XLIFF files into your CAT tool.
- Run the translation against the TM.

You should get exact matching on the plain text and fuzzy matching on the tags. It may be possible to automatically recover 70% of the tags. Depending on the algorithm used to measure quality, this means you will achieve about 80% to 95% matching overall.

- Once the translator has completed the translation, the TM should be exported as a TMX file.

This TMX will correctly tag the DITA block elements as well as correctly segment the sentences, and should therefore be used as the TM for the next DITA-based translation project. For future localization projects, the new TMX should yield exact matching at the segmentation level used for translation (block or sentence).

3.2.1. Advantages of using TMX

Choosing a translation service provider who uses a tool that supports TMX has several advantages. It is possible to migrate your TM between CAT tools that support the industry standard for TM interchange. This is important not only to free you from dependence on a single translation service provider, but also to allow you to fine-tune your segmentation rules to better match your DITA-based XML source documents you'll be sending for translation.

.

3.3. Migrating Legacy TM

This section recommends how best to migrate legacy TM to DITA, for use when translating DITA content moving forward.

When using a CAT tool that does not support TMX, or supports features that would be lost by using TMX, it is recommended to migrate the legacy TM using the tools provided by the CAT tool. When migrating the TM, keep the following points in mind:

- Ensure the TM aligns with the DITA source at the segment level, to achieve the maximum level of exact matching possible for segments.
- Ensure the TM aligns with the DITA inline elements. When migrating to DITA from a non-XML legacy format, formatting tags should be removed, and inline DITA elements should be added. Note that mapping from the legacy inline formatting to DITA inline elements is not always 1:1. Also, DITA has many inline elements that may have no equivalent in the legacy TM.
- Resolve unmatched tags (see Section 3, “Recommended Best Practices”.
- Remove any segmentation rules that are not relevant in the context of DITA. If required, add segmentation rules that apply to the DITA content. Segmentation rules are discussed in Section 3, “Recommended Best Practices”.

4. Notes

- It should be noted that, in general, although sentence level segmentation provides better matching, working with segmentation at the block level improves the quality of the translation. For example, you may need three sentences in Spanish to translate two English sentences. The resulting Spanish translation will read better if the paragraph is translated as a block instead of isolated sentences.
- If the best practices discussed above are used, the first translation of the DITA content can include new content. There is no need to translate the DITA content after migration to DITA before adding new content to the documents.