An OASIS White Paper

Best Practice for Leveraging Legacy Translation Memory when Migrating to DITA

By Gershon Joseph and Rodolfo Raya
For OASIS DITA Translation Subcommittee

27 August 2007

OASIS (Organization for the Advancement of Structured Information Standards) is a not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards. Members themselves set the OASIS technical agenda, using a lightweight, open process expressly designed to promote industry consensus and unite disparate efforts. The consortium produces open standards for Web services, security, e-business, and standardization efforts in the public sector and for application-specific markets. OASIS was founded in 1993. More information can be found on the OASIS website at http://www.oasis-open.org.

The purpose of the OASIS DITA Technical Committee (TC) is to define and maintain the Darwin Information Typing Architecture (DITA) and to promote the use of the architecture for creating standard information types and domain-specific markup vocabularies. The Translation Subcommittee defines best practices and guidelines for DITA authoring, translation and localization, and recommends solutions for industry requirements for consideration by the OASIS DITA TC. The group recommends widespread adoption of these concepts through liaisons with industry, other standards, and providers of commercial and open source tools.

Table of Contents

1. Statement of Problem

Many organizations have previously translated content that was authored in non-XML tools, such as desktop publishing applications. When migrating their legacy content into the new DITA authoring environment, what does an organization do about their legacy translation memory? This legacy translation memory (TM) has been created with large financial investments that can't easily be discarded simply because a new authoring architecture is being adopted.

This paper describes best practices that will help organizations use their legacy TM for future translation projects that are authored in DITA XML. These practices will allow them to minimize the expense of ongoing translations of XML-based content.

In general, there is no need to translate the existing content after migration to DITA before adding new content to the documents. Without following these best practices, the conversion to DITA will be more expensive for each language.

This practice assumes that content reuse in DITA will be based on the use of the conref mechanism rather than on the use of user-defined entities. The discussion about conrefs applies as well to user-defined entities.

2. Recommended Best Practices

This section describes the recommended process at a high level that is independent of tools used and the features they support. This best practice recommends segmenting the TM at the sentence level to achieve better matching to support the migration of content to DITA and prior to translating the DITA content.

It should be noted that, in general, sentence-level segmentation provides better matching. However, working with segmentation at the block or paragraph 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 as isolated sentences. Therefore, you may want to set the TM back to block segmentation following the transform to DITA.

The process includes adjusting the tagging and segmentation rules of your TM so that it is better aligned with the DITA content. 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 is likely to yield substantial translation savings.

Please take time to acquaint yourself with the relevant localization industry standards. Open standards allow you to have more choice and flexibility when establishing the best way to handle the transition to DITA. A full list is provided at the end of this document. At the forefront of these standards are XLIFF for exchanging localization text with language service providers and xml:tm, which takes the reuse principle down to the sentence level and integrates perfectly with DITA.

3. Translation Memory Improvement Process

If you are beginning with non-XML content, most likely in a desktop publishing application, do the following:

1. Migrate the legacy source content to DITA.

2. Update the DITA source with new content as necessary before the document release is to be translated.

3. Using the resources of your TM tools, adjust your legacy TM tagging and the segmentation rules to ensure a high percentage of matches. When tuning your legacy TM, take the following into account:

   Unmatched tags — Unmatched tags can result from conditional text marked up in legacy desktop publishing tools or when block elements contain several sentences that share a common format marker. A number of conditions will generate unmatched tags, including conditional text markers, cross-reference markers, book marks, footnotes, and so on. For example, a paragraph may contain 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 DITA documents. For example, your legacy desktop publishing-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. With DITA, you have a well-defined set of inline or subflows tags which may be very different from the tags used in the legacy format.

Note: Adjusting the legacy TM can be made much easier if you convert the legacy TM to TMX format. TMX is an XML format that makes many of the following editing procedures much easier. If you are using an external localization service provider (LSP) ask if they can provide this service or can give you the TM in TMX format. Your agreement with your LSP should allow you to have access to the TM. Alternatively, you can use tools to align previous source and target language documents at the sentence level and export the result in TMX format. You can then load the TMX into all industry TM products that support TMX

4. Run the DITA content against the modified legacy TM.

   The following points should be kept in mind when translating DITA content:

   If 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.

Ensuring that the TM aligns with the DITA source at the segment level will achieve the maximum level of exact matching possible for segments.

Inline elements may not match at all or may only yield 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 most of the inline tags, which is the main area where matching is prone to fail.

It is often worth the effort to 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 one to one. Also, DITA has many inline elements that may have no equivalent in the legacy TM.

Properly close and/or open unmatched tags in a segment.

Segments without inline elements should yield exact matches from the TM.

The introduction of a conref will disrupt the matching process. However, if a conref refers to an element that contains one or more complete segments and has no unmatched tags or inline elements, the conref text should produce an exact match in the topic in which the conref is resolved.

If a conref is used to replace text within a segment in the source topic, the segment will not be properly matched. For best practices on using conref elements in DITA topics that need to be translated, please see the XREF to CONREF best practice.

Review the segmentation rules defined in the translation tool, and adjust them as necessary to make them relevant in the context of DITA. If required, add segmentation rules that apply to the DITA content. Segmentation rules are discussed above in “Recommended Best Practices.”

5. After the translated content has been approved, the new TM 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 TM should yield exact matches at either the block or the sentence segmentation level.

4. Terms and Acronyms

DITA block element

The contents of a DITA non-inline element, for example <p>, <section> and <table>.

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).

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.

In-context 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 for the interchange of translatable text between CAT tools.

SRX

Segmentation Rules eXchange is an industry standard for establishing and exchanging sentence-level breaks.

xml:tm

XML Based Text Memory is an industry standard that takes the DITA principle of text reuse (both author and translation memory) down to the sentence level. It also provides a standard mechanism for establishing in-context exact matching.