

Proposed Localization Data Technical Committee

The Unicode Consortium is proposing that a new Technical Committee be established to provide profiles of use for sharing data between implementers of localization standards, particularly, XLIFF, TMX, and SRX. This Technical Committee would be governed by the Unicode [Technical Committee procedures](#), which require voting membership in the Consortium for full participation. The procedures would be updated if the new committee is initiated.

Since the first draft of this proposal, the LISA organization has apparently published its intent to shut down its standards work. There is clearly opportunity to engage with those who have been contributing to TMX and SRX to see if the work could be re-hosted to the Unicode Consortium. Where LISA is discussed in this proposal, a new committee would need to instead engage with the former participants in LISA's open standards activities for TMX and SRX.

Technical Background and Proposal

Organizations have long experienced the challenge to provide information to relevant audiences either for commercial reasons or other purposes. With the wide availability of the Internet through various devices, the demand for information in local languages is ever higher. It is clear that in order to meet the volume of demand, more automation and standardization is needed.

1. Purpose

The purpose of the proposed Localization Data Technical Committee is to provide needed *data interchange standards* for localization-related assets. Whether a translation request is completed by human or machine, a few key assets play vital role in the overall process:

- *Translation memory*: a translation memory system stores words or phrases that have been translated previously. The use of translation memory not only ensures the consistency of translated content, and accelerates the speed of translation, but it also reduces the cost of repeat translation requests.
- *Segmentation rules*: segmentation rules define the way to segment text for translation or other text processing. It is used in conjunction with translation memory in order to create memory segments or identify matches within the source content of existing translation memories.

2. Background

The existing localization data standards are unique data standards mainly owned by [OASIS](#) (XLIFF) and [LISA](#) (TMX and SRX). However, there are no specifications on how these standards should be interpreted under interchange scenarios. The owning organizations have not been focused on the development of these data standards in the last few years. Due to the lack of technical leadership, the existing localization standards do not present comprehensive certification requirements. As a result, there is no ability to interchange localization data produced by the industry tools implementing these standards. The current tools on the market

Proposed Localization Data Technical Committee

maintain closed and proprietary data exchange formats while claiming compliance to the existing standards.

3. Benefit

Localization of software information is a key part of the adoption of most software offerings in many countries. The establishment of this workgroup is consistent with the Unicode Consortium's mission statement: "devoted to developing, maintaining, and promoting software internationalization *standards and data*". Currently, multiple organizations from the localization industry are interested in contributing to the further development of these data interchange standards. This proposed committee can provide an engagement opportunity for organizations who have not traditionally participated in Unicode Consortium activities.

4. Scope

The goal is to mature profiles of use for key localization data exchange standards by leveraging the wide reach of Unicode Consortium to form a Localization Data Technical Committee to:

- Gather requirements for extensions of the specified standards.
- Establish extensions or implementations to improve the usefulness of the standards and profiles for interoperability.
- Provide consistent interpretation of the extensions and profiles.
- Focus on a limited set of data standards: XLIFF, TMX, and SRX.

5. Background information on XLIFF, TMX, and SRX

XLIFF

From a Sun Developer Network article

The XLIFF format grew out of a collaboration between a number of companies ... but was soon brought under the management of an OASIS Technical Committee. In April 2002, the first Committee Specification for XLIFF was published. This is available at <http://www.oasis-open.org/committees/xliff/documents/xliff-specification.htm>.

The XLIFF format aims to:

- Separate localizable text from formatting.
- Enable multiple tools to work on source strings and add to the data about the string.
- Store information that is helpful in supporting a localization process.

The XLIFF File:

In its most basic form, the XLIFF file consists of one or more file elements. Each of these contains a header and a body section. The header contains project data, such as contact information, project phases, pointers to reference material, and information on the skeleton file (explained below). The body section contains `trans-unit` elements--the main elements in an XLIFF file.

Proposed Localization Data Technical Committee

The `trans-unit` elements store localizable text and its translations. These elements represent segments (usually sentences in the source file that can be translated reasonably independently). The `trans-unit` elements contain `source`, `target`, `alt-trans`, and a handful of other elements. The example below shows how they would be used.

Example 1. Example of a `trans-unit` Element

```
<trans-unit id="n1">
  <source>This is a sentence.</source>
  <target xml:lang="fr">Translation of "This is a sentence."</target>
  <alt-trans match-quality="100%" tool="TM_System">
    <source>This is a sentence.</source>
    <target xml:lang="fr">TM match for "This is a sentence."</target>
  </alt-trans>
  <alt-trans match-quality="70%" tool="TM_System">
    <source>This is a short sentence.</source>
    <target xml:lang="fr">Fuzzy TM match for "This is a sentence."</target>
  </alt-trans>
</trans-unit>
```

This example shows a pseudo-translated segment. The `trans-unit` element contains an `id` attribute used to determine where the segment goes in the original document. The `trans-unit` element has a `source` and a `target` element as children. The `source` element represents the source text (the text to be translated) in the original document. The `target` element represents the currently accepted translation of the source after linguistic review has taken place.

The example also shows the `alt-trans` elements. These represent translation alternatives for the `source` segment in the `trans-unit` element. A translation alternative is a translation found in a translation memory, a translation generated by a machine translation system, or a translation suggested by a translator or reviewer. These elements contain `source` and `target` elements. In this example, `target` elements are the suggested translations of the `trans-unit` source. The `source` element represents the text that was matched against, from a TM system, for example.

The `alt-trans` element contains attributes such as `match-quality` and `tool`. These provide information about the alternative translations, such as which tool produced them, or in the case of `match-quality`, a measure of the quality of the translation. The algorithm for generating the `match-quality` value in a given `alt-trans` element is specific to the tool that generated it. However, for a translation memory system, it is typically the percentage of words in the `source` element that match the `source` from its database.

TMX

From Wikipedia

TMX (**Translation Memory eXchange**) is an open [XML](#) standard for the exchange of [translation memory](#) data created by [computer-aided translation](#) and localization tools. TMX is developed and maintained by [OSCAR](#) (Open Standards for Container/Content Allowing

Proposed Localization Data Technical Committee

Re-use), a special interest group of [LISA](#) (Localization Industry Standards Association). Being in existence since 1998, the format allows easier exchange of translation memory between tools and/or translators with little or no loss of critical data. The current version is 1.4b - it allows for the recreation of the original source and target documents from the TMX data. TMX 2.0 was released for public comment in March, 2007.

TMX forms part of the Open Architecture for XML Authoring and Localization ([OAXAL](#)) reference architecture.

SRX

From Wikipedia

SRX (Segmentation Rules eXchange) is an [XML](#)-based standard maintained by LISA. It provides a common way to describe how to segment text for translation and other language-related processes. It was created when it was realized that TMX leverage is lower than expected in certain instances due to differences in how tools segment text. SRX is intended to enhance the [TMX standard](#) so that [translation memory](#) (TM) data that is exchanged between applications can be used more effectively. Having the segmentation rules that were used when a TM was created will increase the leverage that can be achieved when deploying the TM data.

Implementation Difficulties: SRX make use of the ICU Regular Expression syntax, but not all programming languages support all ICU expressions, making implementing SRX in some languages difficult or impossible. Java is an example of this.