Open and Flexible Metadata in Localization

How often have you tried to open a Translation Memory (TM) created with one computer-aided translation (CAT) tool in another CAT tool? I assume pretty often. In the worst case, you cannot open the TM. In the best case, you can open it, but data and metadata are lost. You aren't able to tell which strings have been locked, which are under review and so on.

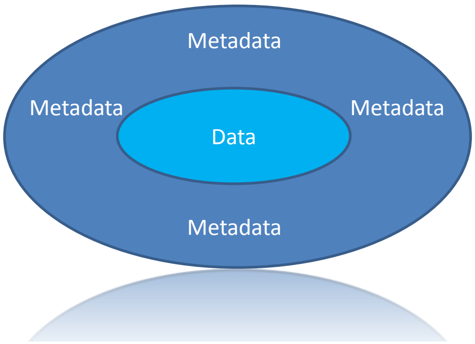
The standard Translation Memory eXchange (TMX), developed by the Localization Industry Standards Association (LISA)'s standards committee called OSCAR (Open Standards for Container/content Allowing Reuse) undoubtedly makes the exchange of TM data easier and does not lock the translators in a specific tool or tool provider. Also, the standard XML Localisation Interchange File Format (XLIFF), developed under the auspices of the Organization for the Advancement of Structured Information Standards (OASIS) is an interchange file format which exchanges localization data and can be used to exchange data between companies, such as a software publisher and a localization vendor, or even between localization tools.

Both TMX and XLIFF are important standards for the localization process. These standards have their own format, though the synergy is there: XLIFF's current version 1.2 borrows from the TMX 1.2 specification, and the inline markup XLIFF support in TMX 2.0 is currently in progress.
There is a range of standard data formats, apart from TMX and XLIFF, such as darwin information typing architecture (DITA), attached to OASIS, Internationalization Tag Set (ITS), put out by W3C, Segmentation Rules eXchange (SRX), affiliated with LISA/OSCAR along with Global Information Management Metrics eXchange-Volume (GMX-V) and so on. Each of these standards focuses on a particular aspect of the localization workflow. For example, DITA is for authoring, ITS for internationalization, SRX for segmentation and GMX-V for counting. There are also some frameworks like the ISO-based Multilingual Information Framework (MLIF) and the Open Architecture for XML Authoring and Localization (OAXAL), which try to take advantage of and combine some of the aforementioned standards.

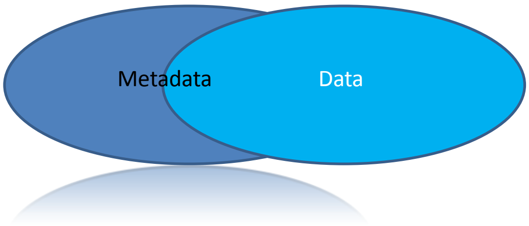
In a period where many CAT tools are created and features are added to their feature lists, what really matters is to have full, proper support of standards by most if not all tools. However, reality proves that currently each CAT tool, commercial or open-source, supports a standard in a different way and each provider customizes it according to their needs, so that in the end the standards cease to have a proper and rigid structure. Supporting standards definitely gives advantage to the tool and tool provider and extends their feature lists, but having poor support is undoubtedly not good to anyone, not the translators, developers or project managers.

In order to check out the adoption of XLIFF by different stakeholders, translators, publishers/clients, tool providers and localization service providers (LSPs), we conducted a survey and asked the respondents for their feedback regarding the general structure of XLIFF, the problems they face using it and any recommendations improving it. Some general comments are that XLIFF can be very complex, and thus simplicity and modularization are needed. Also, clarification of necessary metadata can help to this direction.

Before we go on in detail to XLIFF as a standard to carry much metadata, the distinction between data and metadata is crucial. While data is the actual content, metadata is any data about other data. In other words, metadata keeps data together and connects them in a chain flow. Metadata often surrounds and describes the actual data, like the following diagram shows:



Sometimes though, the metadata is inserted in the actual data and the diagram changes then and looks as follows:



In this case, it is even more important to distinguish between data and metadata, so nothing is lost on the way, data (most importantly) or metadata.

We now take some localization workflow stages: project management, technical writing, translation/localization, generation of content and furnish just some examples of metadata.

- Project management: distinction between translatable and non-translatable files; start and due date of the project; comments; locking strings.
- Technical writing: author’s details (name, e-mail); content domain; style guides used.
- Translation: status of the stings (needs translation, under review, approved/confirmed); context information; linguistic assets used, i.e. TMs, Terminology Databases (TDBs), glossaries or pre-translate with MT; leveraging information (matching percentage).
- Web and software localization: coordinates of menus and dialogs; layout formatting; version control; bug status/report.

The advantages of using metadata are many and diverse. More precisely and in relation to the above, metadata in software localization is important, as most often software versions are based on previous ones. Also, the project manager and the translator benefit from metadata: the project manager using metadata will have a better folder structure and will be able to identify and search the relevant files. The translator is helped by metadata in that he/she sees what to translate and what not, sees the suggested translations provided by TM or MT and so on.

The above stages and metadata are only some examples and can be either increased or decreased depending on the project, process and workflow followed. In most cases, in the localization industry metadata is project-specific, resources-dependent and process-oriented. Generally speaking, metadata provides additional knowledge about digital content, services, processes and the people or organizations who use them; it can be differently applied to any given data depending on the activities and purposes to which it is put.

As aforementioned, metadata is often process or content-focused; that is why metadata in the localization process is needed for various reasons and purposes. Perhaps the most important reason we need metadata is in order not to lose any relevant content. The content should be maintained during the authoring, internationalization and localization process and even after that, for its reuse in future projects.

But having metadata does not solve all these problems. Metadata should be useful. By useful metadata, we mean rich, explicit, easily-transferred metadata. Software developers should write explicit metadata in a form that is understood by those without an engineering background. Also, the metadata should be “understood” and supported by software through the various filters. Then the result is reflected both in terms of cost and time saving.

The question that arises is why should metadata be standardized. Many tool providers may think that standardization restricts their tasks and narrow their focus. Also, they have to either find people who are familiar with working with standards or train their recruited staff. And that is, of course, reflected to time consumption and cost intension. So, the adoption of standards by both clients and vendors is sometimes tricky and it ultimately depends on the motivation of the companies.

The main advantage of standardization of metadata is that there is a common basis of writing and understanding metadata. Standardization is needed to ensure the safe capture, usage and re-usage of metadata, so that actors in a localization process can communicate easily and the localization process is performed in general faster, cheaper and with higher accuracy. Moreover, when standards are open, then more people are involved in projects, more tools are developed and generally there is independence and freedom in selecting the best solution.

XLIFF

XLIFF is an open standard for translation and localization and exchanges data between software publishers, localization vendors or localization tools. It is developed under the auspices of the OASIS Technical Committee, whose purpose is to define and promote the adoption of a specification for the interchange of localizable software-based objects and metadata, so that documents can go through several localization phases. In general, XLIFF supports custom defined workflow metadata during the localization process. During the different stages of localization, the XLIFF document might include data from TM, MT and termbases. Segments can be marked as signed-off, needing review and so on. It also supports metadata such as restrictions on string length, coordinates of dialog boxes, as well as specifying data type for a translation unit.

For example, a simple valid XLIFF example without metadata would include the current version of XLIFF declaration and the elements `file` (including the original file, source and target language and datatype), `body` and translation unit (`trans-unit`), as it is shown in Table 1 below. What the example says is that the US English source word *magazine* is translated into German *Zeitschrift*.

```
<?xml version="1.0" encoding="UTF-8" ?>
<xliff version="1.2" xmlns="urn:oasis:names:tc:xliff:document:1.2"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="urn:oasis:names:tc:xliff:document:1.2
      xliff-core-1.2-transitional.xsd">
       <file original="magazine_without_metadata.txt" source-language="en-US" target-
       language="de-DE" datatype="plaintext">
              <body>
                     <trans-unit id="#1">
                            <source>magazine</source>
                            <target>Zeitschrift</target>
                     </trans-unit>
              </body>
       </file>
</xliff>
```

Example 1. XLIFF example – without metadata

Another example of an XLIFF file, rich in metadata this time, follows in Table 2:

```
<?xml version="1.0" encoding="UTF-8"?>
<xliff version="1.2" xmlns="urn:oasis:names:tc:xliff:document:1.2"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="urn:oasis:names:tc:xliff:document:1.2
      xliff-core-1.2-transitional.xsd">
       <file original="magazine_with_metadata.txt" source-language="en-US" target-
       language="de-DE" datatype="plaintext" tool="TM-ABC">
       <header>
             <phase-group>
                <phase phase-name="review"
                       process-name="Terminology Management"
                       contact-name="Dimitra Anastasiou"
                       contact-email="Dimitra.Anastasiou@ul.ie"/>
             </phase-group>
             <glossary>
               <external-file
               href="file:///C:/MyFolder/MyProject/Multilingual_glossary.htm"/>
             </glossary>
             <note>Please consult the multilingual glossary!
             </note>
       </header>
    <body>
             <trans-unit id="#1">
                    <source>magazine</source>
                    <target>Zeitschrift</target>
                    <alt-trans match-quality="75%">
                           <source>magazine issue</source>
                           <target>Zeitschriftenausgabe</target>
                    </alt-trans>
             </trans-unit>
    </body>
   </file>
</xliff>
```

Example 2. XLIFF example – rich in metadata

In the example in table 2, we have an XLIFF file full of metadata. In `header` element we have metadata about the whole localization process: `phase-name=”review”` and `process-name=”Terminology Management”` as well as the contact details of the person responsible (name and e-mail). We also have a glossary external file and a comment to consult this glossary. Interesting is also the metadata related to the alternative translation element (`alt-trans`), where from the tool TM-ABC we have the entry *magazine issue* translated into *Zeitschriftenausgabe* and also the matching percentage (75%).

Centre for Next Generation Localisation and Metadata

The Centre for Next Generation Localisation (CNGL) is an academia-industry partnership project that integrates MT technology, speech-based interfaces and personalization, multilingual digital content management and localization workflows. With more than 100 researchers working in the terms of CNGL and covering the whole localization process, a common open metadata model is necessary for the better collaboration both for people (such as developers, translators, communities) and processes (authoring, localization, maintenance and management).

A metadata group is set up within CNGL; one of its tasks currently is to check the types of metadata models utilized within CNGL and foresee which extra metadata is needed in the future. This future metadata can be accordingly recommended for the next XLIFF releases. As for our initial recommendations, tools should have converters to be interoperable. What we need is open exchange of workflow data and to build global concern for standardization. This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at the University of Limerick.