Survey on the Use of XLIFF in Localisation Industry and Academia

Dimitra Anastasiou

Centre for Next Generation Localisation,

Localisation Research Centre,

Department of Computer Science and Information Systems

University of Limerick,

E-mail: Dimitra.Anastasiou@ul.ie

Abstract

The XML Localisation Interchange File Format (XLIFF), developed currently under the auspices of the Organization for the Advancement of Structured Information Standards (OASIS), first released in April 2001 (v1.0), is an open standard for translation and localisation which allows for the exchange of data between software publishers, localisation vendors, or localisation tools. The OASIS XLIFF Technical Committee defines and promotes the adoption of a specification for the interchange of localisable software-based objects and metadata, so that the real content (data) and the data about data (metadata) can be transmitted smoothly through several localisation phases, from digital content creation, through to internationalisation, localisation, and actual content generation. This paper presents the results of an XLIFF survey that was conducted in order to evaluate the adoption of XLIFF by different stakeholders, both from industry and academia, be they tool developers, customers, or translators. The general structure of XLIFF was rated as good by more than half of the respondents, the synergy with other standards, such as ITS (W3C) and TMX (LISA) was rated more than desirable, while some of the changes recommended for XLIFF's structure are increased simplicity, modularisation, clarification of necessary metadata, and better workflow control.

1. Introduction

According to the Localisation Industry Standards Association[1] (LISA), localisation is defined as follows:

> *"Localisation involves the adaptation of any aspect of a product or service that is needed for a product to be sold or used in another market."*[2]

Localisation is distinct from translation, because it is not only text being transferred from one source language (SL) to a target language (TL), but also, icons/images, audio, video, colours, layout, and other "aspects of products or services".

In terms of localisation, while *data* is the actual content to be localised, *metadata* is the data about data, i.e. describing, explaining, and processing other data. Metadata is undoubtedly as important as data, since the former provides structure and order to the latter, and generally defines a clear workflow.

Localisation metadata not only defines and supports a clear workflow, but also connects the data present at different localisation workflow stages. Metadata is very useful during the digital content creation, linguistic annotation, content maintenance, translation (Translation Memory (TM) and Machine Translation (MT) usage), proofreading/postediting, content generation of the localised content, and process management in general.

In section 2 of this paper we examine some localisation standards, and focus on the XML Localisation Interchange File Format (XLIFF) standard. Section 3, the main body of the article, focuses on a survey that we conducted pertaining to the adoption of XLIFF, the difficulties of supporting it, recommended changes, etc. In section 4 we introduce the Centre for Next Generation Localisation (CNGL) project and we describe the tasks of the metadata group set up within CNGL. A conclusion and future prospects of our research combining our membership of the XLIFF Technical Committee (TC) and chairing the metadata group in CNGL are found in the last section of the paper 5.

2. Standards – Related work

There are many standards bodies, such as ISO, W3C, LISA, OASIS, etc. Each of these organizations manages standards on different aspects of information management and technology topics; we focus though more on the localisation-related standards.

In 2006, an initiative called MultiLingual Information Framework (MLIF)[3] was standardised (ISO/TC37/SC4) providing a common platform for all existing tools and promoting the use of a common framework for the future development of several different formats: TMX (LISA), XLIFF (OASIS), etc. MLIF introduces a metamodel for multilingual content in combination with data categories as a means to ensure interoperability between several multilingual applications and corpora. MLIF examines at morphological description, syntactical annotation, or terminological description. An important point about MLIF is that it does not propose a closed list of description features, but rather provides a list of data categories, which are much easier to update and extend.

In addition, the Open Architecture for XML Authoring and Localization[4] (OAXAL) is another framework which takes advantage of the Darwin Information Typing Architecture (DITA) standard from OASIS and also

[1] http://www.lisa.org/, 29.04.2010

[2] http://www.lisa.org/Localization.61.0.html, 29.04.2010

[3] http://mlif.loria.fr/ and http://www.tc37sc4.org/new_doc/ISO_TC37-4_N266_WD_Multilingual_resource_management.pdf, 29.04.2010

[4] http://www.xml.com/pub/a/2007/02/21/oaxal-open-architecture-for-xml-authoring-and-localization.html, 29.04.2010

XML-based text memory (xml:tm) standard from LISA's standards committee called Open Standards for Container/content Allowing Reuse (OSCAR). It is important to point out that XLIFF is an OASIS standard, while OAXAL a TC specification.
A localisation standard released by W3C, the Internationalization Tag Set[5] (ITS), is designed to be used by schema developers, content authors, and localisation engineers to support the internationalisation and localisation of schemas and documents. It includes data categories and their implementation as a set of elements and attributes. In ITS there is a global and local approach; one of the benefits of the global approach is that the content authors make changes in a single location, rather than by searching and modifying the markup throughout a document.
To give an example related to localisation in ITS global approach, one can look at the translateRule element. It includes a translate attribute with boolean value (in this case "no") and a selector. The selector contains an XPath expression which selects the nodes. Rules apply to these nodes (see Table 1).

```
<its:rules
xmlns:its="http://www.w3.org/2005/11/its"
 version="1.0">
<its:translateRule translate="no" selector
="//code"/>
</its:rules>
```

Table 1: ITS example of global approach
Source: http://www.w3.org/TR/its/#translatability-implementation

The conversion tool from ITS2XLIFF[6] (v 0.6) developed by Felix Sasaki[7] is worth mentioning here. His tool allows users to generate up to date XLIFF files (v 1.2) from XML files for which W3C ITS rules are available.
LISA/OSCAR standards[8] include:

- Translation Memory eXchange (TMX);
- Segmentation Rules eXchange (SRX);
- Term-Base eXchange (TBX);
- XML Text Memory (xml:tm);
- Global Information Management Metrics eXchange - Volume (GMX-V).

TMX is probably the most well known LISA/OSCAR standard as it exchanges TM data between applications, being commercial or open-source.
We now turn our focus to XLIFF. For information about a brief history see Reynolds and Jewtushenko (2007). As previously mentioned, XLIFF is managed by OASIS, a not-for-profit consortium which also produces many Web service procedures. According to OASIS, XLIFF is defined as follows:

> *"XLIFF is […] designed by a group of software providers, localisation service providers, and localisation tools providers. It is intended to give any software provider a single interchange file format that can be understood by any localisation provider."*[9]

An example [10] of an XLIFF translation unit `<trans-unit>` element is visible in Table 2:

```
<trans-unit id="1" restype="button"
resname="btnSAVE"
coord="12;124;16;80">
<source xml:lang="en-us">Save</source>
<target xml:lang="de-de"
state="needs-translation">Speichern
</target>
</trans-unit>
```

Table 2: XLIFF example of *trans-unit* element

The actual data in this example shows that the button's name *Save* in the source language (SL) English (US) English means *Speichern* in German. XLIFF is a standard that can carry a large amount of metadata, as we can see from the example in Table 2: resource type (`restype="button"`), coordinates (`coord="-12;124;16;80"`), and translation state (`state="needs-translation"`).
Another XLIFF element important to localisation is the `alttrans` element (see Table 3). This element contains possible alternative translations, e.g. in `<target>` elements along with optional context, notes, etc. (see Table 3):

```
<alt-trans match-quality="80%" tool="XYZ">
<source>Save as</source>
<target xml:lang="de-DE"
phase-name="pre-trans#1">Speichern unter
</target>
</alt-trans>
```

Table 3: XLIFF example of *alt-trans* element

Here we have a matching percentage of 80%, because in the TM of the tool XYZ we had the entry *Save as* translated as *Speichern unter*.
There is a relationship between TMX and XLIFF in that XLIFF 1.2 borrows from the TMX 1.2 specification, but they are different standards, each having their own format. Inline markup XLIFF support in TMX 2.0 is currently in progress.
Based on Rodolfo Raya's (2007) article "*XML in localisation: Reuse translations with TM and TMX*", we created the following table which distinguishes between TMX and XLIFF and also brings them into symbiotic relationship, see Table 4.

	Definition	Synergy	Conversion
TMX	Standard for the exchange of TM data	Should be used as a complement to	XSL transformation to convert an

[5] http://www.w3.org/TR/its/, 29.04.2010
[6] http://fabday.fh-potsdam.de/~sasaki/its/, 29.04.2010
[7] Also, Christian Lieske contributed to the development of this tool.
[8] http://www.lisa.org/OSCAR-LISA-s-Standa.79.0.html, 29.04.2010
[9] XLIFF Specification 1.2: http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html, 29.04.2010
[10] This is not a full valid XLIFF file, but only an element and its contents.

	created by CAT and localisation tools	XLIFF	XLIFF file to TMX format
XLIFF	Format for exchanging localisation data	Possible translations contained in the `<alt-trans>` elements of an XLIFF file are extracted from a TM database	

Table 4: XLIFF-TMX

It is noteworthy that XLIFF has a working relationship with LISA/OSCAR standards and is a requirement for the standards TMX, GMX-V, and xml:tm.

1. XLIFF survey

As aforementioned, XLIFF was first released in 2002 and its latest version is 1.2 (approved as a Specification in February 2008). Currently, the XLIFF TC is working on the specifications of the next XLIFF release 2.0[11].

Recently both commercial and open-source tools have supported XLIFF. Examples of commercial tools supporting XLIFF are Swordfish, XTM, SDL TRADOS, Alchemy Catalyst, memoQ, and some open-source examples are OmegaT, Virtaal, etc. There are more tools which support the format of XLIFF, but mentioning these tools and their diverse support is outside the scope of this particular paper.

In fact, the interest in XLIFF is not total and it is only in recent years that more and more tools have begun to support it. But is this XLIFF support full and proper support or does it only cover the basic features? Also, how often are there cross-tool operations and when are they successful?

These reasons motivated us to conduct a survey about XLIFF in terms of primary research. One questionnaire consists of eight questions, five of which are multiple choice questions and three ask for general feedback. The survey was created by the author with the help of Reinhard Schäler, director of the Localisation Research Centre (LRC). The questionnaire is available online[12] and copies of the questions can be found in the Appendix (section 8).

70 respondents completed the questionnaire. Half of the responses were received after distributing the questionnaire at the tc world conference in Wiesbaden, 2009; the other half was collected by sending the questionnaire to mailing lists.

The 70 respondents of the survey were either from industry (tool providers (33%) and localisation service providers (17%)) or from Academia (CNGL[13] researchers (22%)). The remainder were translators (11%), content publishers (6%), and others, e.g. consultants, students (11%). The distribution of the survey's respondents (first question of the questionnaire) can be seen in Diagram 1:

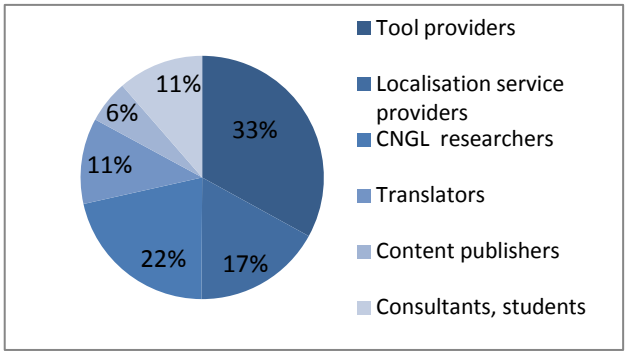

Diagram 1: Respondents

The second question posed is whether the technologies and tools the respondents use are XLIFF-compliant. This question received a positive response from 33% and a negative response from 20%. The remaining 47% was split to four categories. The first category is by those where some technologies are XLIFF-compliant while some others are not (20%). The second category featured17% who heard of but were not exactly aware of what XLIFF is about, and the third category concerns 3% who gave other answers, such as "not yet, as tools are now compliant". In the fourth category, 6% stated that they had never heard of XLIFF before. The distribution of the percentages regarding this question is shown in Diagram 2:

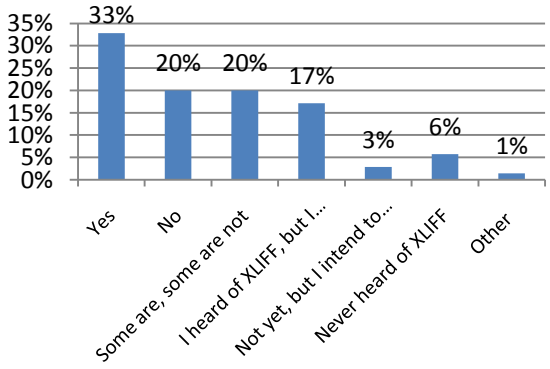

Diagram 2: XLIFF-compliance

The third question concerned the use of XML in XLIFF and whether it should be supported by namespaces. This question is included, because, in our opinion, tool providers often customise the XLIFF document at an extreme level. The responses were diverse; some people would prefer for more extensibility, while some others argue that XLIFF is already too flexible. What most respondents answered (and we agree with) is that if XLIFF is extremely user-defined and there are custom namespaces for every different CAT tool, then the cross-tool operations will lead to data loss. According to some respondents' answers, the solutions to that would be "stronger standards and not just guidelines", "tools that comply to proper XLIFF coding", and "starting with a simple base and expanding".

Moving towards the fourth question "Should there be more synergy between XLIFF and other standards?", the predominant answer is yes. The feedback received was that XLIFF should be in synergy with TMX, ITS, and GMX-V. According to some respondents, the

[11] http://wiki.oasis-open.org/xliff/XLIFF2.0/FeatureTracking, 29.04.2010

[12] http://ai.cs.uni-sb.de/~stahl/d-anastasiou/Survey/XLIFF_questionnaire.pdf, 29.04.2010

[13] http://www.cngl.ie/, 29.04.2010

LISA/OSCAR process of developing specification should be more open, so that LISA and OASIS work better together. One noteworthy answer was that strict synchronisation is not necessary, as every standard focuses on a particular part of the localisation workflow.
The fifth question asked about problems that users of XLIFF face. If we categorise the answers, we come up with the following:

- Difficulties converting end client formats;
- Tools unable to handle and support XLIFF (and also the same way);
- Infrequent use by publishers/clients and lack of acceptance by professional translators;
- Lack of filters.

Based on these problems, we asked the respondents what they would change in XLIFF. Again, we tried to categorise the feedback which is as follows:

- Simplification of the inline markup;
- Stronger compliance requirements ;
- Clarification of necessary metadata;
- Better workflow control;
- Support for multilingual content.
- Tools which create and use rather than just import XLIFF;

We ended the questionnaire by asking how people would evaluate the general structure of XLIFF. Most regarded it as good (59%), followed by very good (24%), and average (17%); nobody chose the "not good" option.

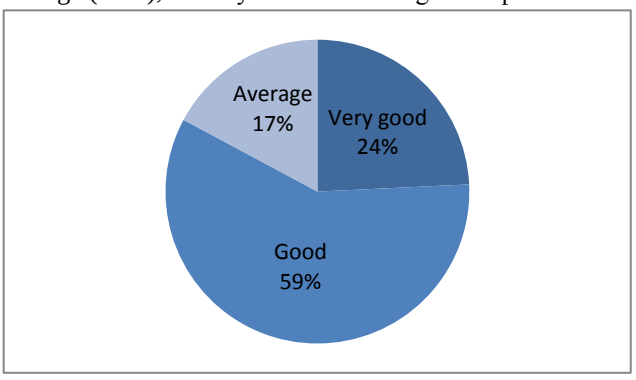

Diagram 3: General structure of XLIFF

2. CNGL

The Centre for Next Generation Localisation (CNGL) project is an Academia-Industry partnership with 100 researchers working on MT and Speech, Digital Content Management, Next Generation Localisation, and System Framework.
As the chair of the metadata group set up in June 2009 in the terms of the CNGL project, the author investigates which metadata is currently used in this project and makes recommendations for future metadata requirements. The goal of the metadata group is to develop a framework which subsumes all of the metadata which must ensure the integrity and interoperability of data as it passes through the areas of content production, localisation, and consumption, as well as asset and process management.
As a member of the XLIFF Technical Committee, the author also examines the use of XLIFF within CNGL; more precisely, we see whether XLIFF's specifications suffice for the CNGL's needs and if not, we collect XLIFF's limitations and make recommendations for the next XLIFF releases.

3. Conclusion and Future prospects

A clear distinction between data and metadata is necessary, particularly in the process of developing specification for standards. Our survey has shown that XLIFF's structure is generally regarded as good, although more simplification, modularisation, as well as more and better adoption by both tools and customers is required. All of the feedback was useful and certainly the XLIFF TC takes that on board and will go towards direction that suffices the needs of the users.
Our future work will be divided in two directions. Firstly, we intend to provide a common metadata framework within CNGL which subsumes all meta-information needed at the different localisation stages. Secondly, we started collecting and arranging, in an hierarchical order, the metadata that exists in XLIFF v1.2 and make recommendations for the next release. To sum up, as chair of the metadata group in CNGL and a member of the XLIFF TC, the author intends to take the outcomes of CNGL research and implement it into the XLIFF standard.

4. Acknowledgements

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.

5. References

Raya, R. (2007). XML in localisation: Reuse translations with TM and TMX. http://www.maxprograms.com/articles/tmx.html and http://www.ibm.com/developerworks/library/x-localis3/.

Reynolds, P., Jewtushenko, T. (2005). What Is XLIFF and Why Should I Use It?. *XML journal:* http://xml.sys-con.com/node/121957?page=0,3.

XML Localisation Interchange File Format (XLIFF) 1.2 Specification: http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html

6. Appendix

Questions of the XLIFF questionnaire

1. You are a: Tools provider/Content Publisher/LSP/Translator/Consultant/Other
2. Are the technologies and tools you use XLIFF-compliant?
3. If you have not implemented XLIFF, why not?
4. What is your opinion about the XML implementation in XLIFF, e.g. namespaces?
5. Should there be more synergy between XLIFF and other standards (Internationalization Tag Set – w3c or LISA' standards). In which way?
6. Which problems do you face using XLIFF?
7. What changes would you recommend in XLIFF?
8. What is your opinion about the general structure of XLIFF?