Localisation Standards and Metadata

Dimitra Anastasiou and Lucia Morado Vázquez

Centre for Next Generation Localisation,
Localisation Research Centre,
Department of Computer Science and Information Systems
University of Limerick, Ireland
{Dimitra.Anastasiou,Lucia.Morado}@ul.ie

Abstract. In this paper we describe a localisation process and focus on localisation standards. Localisation standards provide a common framework for localisers, including authors, translators, engineers, and publishers. Standards with rich semantic metadata generally facilitate, accelerate, and improve the localisation process. We focus particularly on the XML Localisation Interchange File Format (XLIFF), and present our experiment and results. An html file after converted into XLIFF, travels through different commercial localisation tools, and as a result, data as well as metadata are stripped away. Interoperability between file formats and application is a key issue for localisation and thus we stress how this can be achieved.

Keywords: experiment, interoperability, localisation, metadata, standards, standardisation, XLIFF.

1 Introduction

This paper describes the localisation process and focuses on interoperability issues which are facilitated through standardised metadata.

According to Schäler [1], "localisation is the linguistic and cultural adaptation of digital content to the requirements and locale of a foreign market, and the provision of services and technologies for the management of multilingualism across the digital global information flow". Multilingualism is achieved through localisation and more precisely its vital part which is translation, while the global information flow refers to localisation workflows. In this paper we focus on the interoperability of tools and translators in a localisation workflow, and stress the importance of metadata and standards.

Translators in the localisation industry are currently very heavily dependent on tool and tool providers. Translators often receive projects because they have acquired (most often purchased) and use a specific tool and conversely are not chosen for other projects, if they do not have the required tool. This kind of lock-in is characteristic of mainstream localisation, and does not provide equal opportunities to all translators. Open source translation tools and also open standards have only started to appear in recent years in order to achieve, among others, independence and equality.

S. Sánchez-Alonso and I.N. Athanasiadis (Eds.): MTSR 2010, CCIS 108, pp. 255–274, 2010.

There are many localisation tools today, both proprietary and open source with many advanced features related to translation, such as resources/references lookup, project management, etc. These tools and components have to talk to each other in order to achieve interoperability. Thus file formats should be flexible and for this reason, open standards are needed. When these open standards are enriched with explicit semantic metadata, the whole translation and localisation process is facilitated and accelerated.

In section 2 we introduce localisation, and in subsections 2.1 and 2.2 we define localisation metadata and present some localisation standards respectively. We focus particularly on the XML Localisation Interchange File Format (XLIFF) as a standard which carries much metadata (section 3). Both the positive and negative sides of standardisation in general, and of XLIFF in particular, are presented in section 3.1. Section 4 focuses on an experiment that we carried out to check the interoperability of standards and their support by a combination localisation tools. We finish with a conclusion and future prospects in section 5.

2 Localisation

In this section we present localisation in general, its approaches, and describe the digital multilingual revolution we are currently experiencing.

Localisation is the process of adaptation of digital content to a target audience; digital content is not only text in digital form, but also multimedia. There are different conceptualisations of people and images across different languages and cultures throughout the world, and localisation tries to capture these different aspects and transform them from one into another. A combination of language and culture is called *locale* and is characterised by unique linguistic, cultural, and physical as well as technical elements. A language or a culture alone cannot form a locale, only a combination of them. Some examples of locale differences can be found in Anastasiou & Schäler [2]. Moreover, in his general concept of personalisation, Wade [3] supports a groundbreaking idea of a "locale of one", meaning that a single person has his/her specific needs, abilities, knowledge, preferences, choices, etc.

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

To start with, usage and sale of products in distinct markets requires access to these markets. When everything is proprietary, from tools to file formats, then this access is limited. This limited access is a common characteristic of ordinary or so-called mainstream localisation which leads to a close and protected relationship between various stakeholders, such as customers-vendors, including publishers, project managers, engineers, testers, translators, proofreaders, and so on. In addition, mainstream localisation is business-oriented, and driven by short-term return on investment (ROI). In out-of-ordinary localisation, access is supported by open standards, open source tools, and interoperable file formats. Furthermore, the relationship between localisers is open and straightforward.

[1] http://www.lisa.org/, 10/06/10.
[2] http://www.lisa.org/Localisation.61.0.html, 10/06/10.

Today we live in a world where digital content is abundant, as it can be developed easily and quickly, and by active users without advanced engineering background (user-generated content (UGC) in social media). Users today request digital content, and enterprises respond to this request. Searching and retrieving information in the users' native language is not any longer a wish, but a requirement. Users select, comment, and generally have a say and participate actively in the whole process. As competition increases, more and more enterprises offer their documentation, websites, and user assistance in more languages.

Undoubtedly, the development of digital content has some disadvantages which should not be ignored. Its management becomes complicated, time-consuming, and cost-intensive. Very often, although the data exists, it is difficult or impossible to identify it. Content Management Systems (CMSs), web crawlers, and information retrieval (IR) tools are tools and technologies which help manage, organise, and find data. Searching is vital and yet often it is not available everywhere or badly implemented (e.g. in wikis, forums, and generally social networking sites). Search Engine Optimization (SEO) is one of the fields which have made interesting progress towards the use of metadata (or meta elements in this field specifically). To sum up, the higher the volume of digital content, the more difficult to manage it; this is the reason why metadata is needed. There are different levels of modularisation and granularity of metadata, but simplified metadata, such as author name, creation date, process status, etc. help digital content management. In the next section we provide more information about metadata and localisation metadata in particular.

2.1 Metadata

In this section we describe what metadata is, particularly pertaining to localisation processes. Metadata is generally defined by a high percentage of literature as "data about data". Metadata became important in the beginning of 1990s with the appearance of the Web and in parallel, the need to scale and filter information. Tim Berners Lee, the inventor of the Web, defined metadata in 1997 as "machine understandable information for the Web". The first metadata workshop took place in Dublin, Ohio and created the "Dublin Core Metadata" set of elements. The Dublin Core Metadata Initiative[3] (DCMI) is today an open organisation engaged in the development of interoperable metadata standards that support a broad range of purposes and business models.

Later in 2004 the Resource Description Framework (RDF) was published as a language for representing and linking metadata about resources on the Web, such as the title, author, and modification date of a webpage, copyright and licensing information about a Web document, or the availability schedule for a shared resource. It is important to note that RDF can also be used to represent information about things that can be identified on the Web, even when they cannot be retrieved.

Interoperability between standards is crucial, as some applications may read some schemas[4], while some others may not. Filters/converters which translate one schema to another do exist, often only internally to the application itself though. Recently,

[3] http://dublincore.org/, 10/06/10.

[4] A schema is a model for describing the structure of information.

DCMI Metadata Terms were published as an RDF schema[5] to which term URIs are redirected and the vocabulary management tool has now been published as an open source project.

In the context of localisation, today data is very often lost when a user uses a new version of a localisation tool or after converting one format to another. Mistakes like sending comments for translation or displaying the wrong status of the strings happen on a daily basis. An experiment transferring a file through different localisation tools is described in section 4.

Metadata is needed in localisation to identify and search translatable and untranslatable resources having a metadata-defined folder structure, provide the translatable files to the suitable translators, identify which linguistic resources have been used, which status a translation unit has, etc. Localisation tools should use interoperable standards, so that the content is displayed correctly, regardless of the file imported. Visualisation of metadata, e.g. a lock for locking strings or an eye icon next to the strings to be reviewed, facilitates and accelerates the localisation process. To be more precise, in software localisation, the version management metadata is important, as most often software versions are based on previous ones.

In general, localisation includes a complex workflow, covering authoring, linguistic annotation and language processing, translation, maintenance of translated content (through translation memories), and generation of target text. Metadata connects data existing at all these and even more stages. In parallel, metadata has to be open and flexible, so that the access is not limited, and semantically rich, so that it is understandable by all people and applications involved. We summarise the above paragraphs in our description of localisation metadata:

> *"Localisation Metadata connects the data present at different stages of the localisation process, from digital content creation, annotation, and maintenance, to content generation, and process management. The usage of open, rich, and flexible metadata is an effective step towards the aggregation and sharing of data between localisation sub-processes."*

As in many other fields, it is important to answer who writes metadata and for whom, why, what kind of metadata, and when. To give an example, a *project manager* should lock some *strings* which should not be translated *before* sending that to a *translator*. This reduces the time to market (TTM), is economical, and leads to a higher quality translation output.

Although (localisation) metadata is important and has obvious advantages, often it is project, process, and workflow-dependent. Depending on the resources of a project and/or of a localisation company, metadata can be adapted and tailored to specific needs. Generally speaking, metadata provides additional knowledge about digital content, services, processes, and the people or organisations who use them; it can be applied to data in different ways depending on the activities and purposes to which it

[5] http://dublincore.org/2008/01/14/dcterms.rdf, 10/06/10.

is put. Generally, the more resources available, the more metadata is needed; also, the more complicated workflow, the higher the granularity of metadata.

As for the reasons why metadata is needed, the most important is in order to maintain the content as it is transferred via different people and applications. The primary goal is that data, and the secondary that metadata should not be lost at any stage of a localisation workflow; but if the latter is well structured and managed, the former can be easily achieved. Data would be, for example, the actual strings to translate, while metadata could refer to the status of these strings. Cost and time savings are an immediate result of metadata management. The whole process is accelerated, because, for example, communication details between stakeholders (e.g. project managers and translators' names) are easily captured in metadata.

To sum up, metadata follows a cycle, and is constantly being adapted, customised, and tailored to the needs (staff, money, and technology resources) of each organisation. To have a common basis for creating, reading, understanding, searching, retrieving, and transferring metadata, metadata should be standardised. Side by side with standardised metadata, data is safely captured, used, and re-used. A standardisation process can take a couple of years, until the specifications become valid and accepted both by a dedicated committee and a community. The adoption of standards takes even longer, as many people believe that standardisation restricts their scope and have to fit in a box that other people want to control. In fact, in the beginning, it is difficult to translate one format into another and generally properly handle a new format, but it is undoubtedly, worthy in order to have easy data loss transfer through various people and applications. We believe that standardisation connects people and technologies and the next subsection tries to highlight this aspect.

2.2 Standards

Standards can vary from web services to telecommunication standards; there are so many standards and developing organisations, and thus a full coverage of all these is outside the scope of this paper. Some standards developing organisations are ISO (International Organization for Standardization), IEEE (Institute of Electrical and Electronics Engineers), W3C (World Wide Web Consortium), LISA (Localization Industry Standards Association), and OASIS (Organization for the Advancement of Structured Information Standards); here we focus on the translation and localisation standards developed by the latter three organisations.

As far as the definition of a standard is concerned, ISO/IEC Guide 2:2004 [4] provides the following general definition of a standard:

> *"[Standard is]a document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context".*

To create a standard, recognised and approved members from industry and academia join a committee and vote regularly for specifications, rules, and guidelines

which characterise the standard. According to ISO standardisation process, there are six stages: proposal, preparatory, committee, enquiry, approval, and publication stage. More information can be found under development process[6] of ISO processes and procedures site.

The standards we are going to describe now follow the order of a localisation workflow, i.e. authoring, internationalisation, and localisation, as is depicted in the following diagram:

Fig. 1. Authoring, internationalisation, and localisation standards

At authoring stage, the standard which stands out is DITA (Darwin Information Typing Architecture) managed by OASIS. ITS (Internationalization Tag Set) is the internationalisation standard put out by W3C, while XLIFF is a standard carrying localisation data and metadata, and is under the auspices of OASIS. XLIFF will be briefly described in this section, but more XLIFF details and examples will follow in section 3.

DITA is an XML-based specification which processes modular and extensible topic-based information types as specialisations of existing types. It allows for the introduction of specific semantics for specific purposes. An example[7] of DITA (1) follows:

(1) DITA example

```
<task id="installstorage">
 <title>Installing a hard drive</title>
   <shortdesc>You open the box and insert the drive.
   </shortdesc>
 <prolog><metadata>
   <audience type="administrator"/>
  <keywords>
      <indexterm>hard drive</indexterm>
      <indexterm>disk drive</indexterm>
  </keywords>
 </metadata></prolog>
<taskbody>
 <steps>
```

[6] http://www.iso.org/iso/standards_development/processes_and_procedures/how_are_standards_developed.htm, 10/06/10.

[7] The example is retrieved from DITA OASIS online community portal: http://dita.xml.org/what-do-info-typed-dita-topic-examples-look, 10/06/10.

```
  <step><cmd>Unscrew the cover.</cmd>
   <stepresult>The drive bay is exposed.</stepresult>
  </step>
  <step><cmd>Insert the drive into the drive bay.</cmd>
   <info>If you feel resistance,try another angle.</info>
  </step>
</steps>
</taskbody>
 <related-links>
      <link href="formatstorage.dita"/>
      <link href="installmemory.dita"/>
 </related-links>
</task>
```

As we see in the DITA example, there is a metadata tag <metadata>, i.e. explicit statement, including information such as keywords (index terms) and audience type. Then the steps and the steps' results are provided. In addition, external links to other related DITA files are available.

After the authoring stage, comes internationalisation which involves the isolation of language and cultural conventions, so that the product does not need any technical engineering after it is sent to be localised. Here we refer to the ITS standard developed by W3C. Directionality (right-to left and left-to-right text/override), which is very important for internationalisation, is supported by the ITS specification. Also, localisation notes <locNote> can be made including their types and/or URI references. The ITS specification consists of so-called "data categories", a set of elements and attributes. Perhaps the most important data category is <translate>, as it expresses information about whether the content of an element or attribute should be translated or not. The values of this data category are "yes" (translatable) or "no" (not translatable). The selection of XML node(s) may be explicit (using local approach) or implicit (using global rules). These approaches can be compared to the style attribute and the style element in HTML/XHTML, respectively. An ITS example[8] (local approach) follows below:

(2) ITS example

```
<dbk:article
xmlns:its="http://www.w3.org/2005/11/its"
xmlns:dbk="http://docbook.org/ns/docbook"
its:version="1.0" version="5.0" xml:lang="en">
  <dbk:info>
      <dbk:title>An example article</dbk:title>
      <dbk:author its:translate="no">
        <dbk:personname>
            <dbk:firstname>John</dbk:firstname>
            <dbk:surname>Doe</dbk:surname>
```

[8] http://www.w3.org/TR/its/, 10/06/10.

```
        </dbk:personname>
        <dbk:affiliation>
            <dbk:address>
                  <dbk:email>foo@example.com</dbk:email>
            </dbk:address>
        </dbk:affiliation>
      </dbk:author>
  </dbk:info>
<dbk:para>This is a short article.</dbk:para>
</dbk:article>
```

Here we see that we do not have an explicit metadata tag, but metadata is still available, such as title, author, person's name (first name and surname), affiliation, address, and contact e-mail. We also see that that all author's information should not be translated as indicated by the data category translate="no": `<dbk:author its:translate="no">`.

After internationalisation, translation and localisation are the processes that involve actual adaptation of content from a source to a target language. We refer to LISA's standards, ISO's MLIF (MultiLingual Information Framework), and also OASIS's XLIFF standard.

Starting with LISA, its dedicated standards committee called OSCAR[9] (Open Standards for Container/content Allowing Reuse) developed and manages the following standards:

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

TMX, maybe the most well known translation standard makes the exchange of translation memory (TM) data between applications possible. TBX is also for exchange, but of data coming from terminological databases. TBX includes two modules: a core structure, and a formalism for identifying a set of data-categories and their constraints, both expressed in XML. It is noteworthy that TBX is based on the ISO standard 30042. More precisely, the TBX framework defined by ISO 30042:2008 supports analysis, descriptive representation, dissemination, and interchange of terminological data in various computer environments[10].

Coming back to TMX, after realising that its leverage was not high, because every tool segmented the text to be translated in a different way, SRX was developed to describe how translation and other language-processing tools segment text for processing. Thus the parallel implementation of TMX with SRX increases the TMX leverage, as it allows for the transmission of the segmentation rules that were used when a TM was created.

[9] http://www.lisa.org/OSCAR-LISA-s-Standa.79.0.html, 10/06/10.
[10] http://www.iso.org/iso/catalogue_detail.htm?csnumber=45797, 10/06/10.

As far as GMX-V is concerned, it quantifies the workload for a given localisation or translation task, not only by word and character count, but also by counting exact and fuzzy matches, (alpha)numeric text units, standalone punctuation, etc. The XML customisation allows for page counts, file counts, screen shot counts, etc., which are useful for localisation processes. It should be noted that GMX-V is based on the following well-defined standards: XLIFF, Unicode ISO 10646, and Unicode TR29-9. In addition, there are other components of GMX, i.e. complexity (C) and quality (Q) (both are proposed, but not yet defined). More information about GMX can be found in Zydroń [5].

As for the xml:tm standard, it allows text memory including translations to be embedded within XML documents. This standard is mainly used in the Open Architecture for XML Authoring and Localization[11] (OAXAL) reference model which is related to the authoring and translation aspects of XML publishing. OAXAL encompasses the following open standards, apart from xml:tm: XML, Unicode, ITS, SRX, GMX, TMX, Unicode TR29, XLIFF, and open standard XML vocabularies, such as DITA.

The last LISA standard to be briefly described is Term Link; it is proposed, but is not an official standard yet. It is an XML namespace-based notation that enables identified terms within an XML document to be linked to an external XML termbase, including those in TBX format. The main benefit of Term Link is that users have access to terminological data stored remotely.

We will now look at another standard related to localisation: MLIF (ISO CD24616). According to Cruz-Lara et al. [6], MLIF does not only describe the basic linguistic elements (sentence, syntactic component, word, part-of-speech), but also represents the structure of the document (title, paragraph, section). MLIF proposes a metamodel and data categories to represent and exchange multilingual textual information. The authors describe the TMX-MLIF interaction in their paper (p. 55), which includes the stages extraction, translation, and merging. The extraction process gives i) a "Skeleton File" containing all translation memory (TM) formatting information and ii) an MLIF file in which only relevant linguistic information is stored. Then an XSL style-sheet could transform an MLIF document into a TMX document (which does not contain any formatting information). Once the translator translates the source text, another XSL style-sheet allows the transformation of a TMX document back into MLIF. Finally, the new MLIF document (containing the translation) is merged with the "Skeleton File" in order to obtain a new TMX formatted document.

We now focus on the XLIFF standard, which is developed and managed by OASIS. XLIFF is an interchange file format which exchanges localisation data and can be used to exchange data between companies, such as a software publisher and a localisation vendor and between localisation tools.

We recently conducted a survey (see Anastasiou [7]) to check the adoption rate of XLIFF by different people in localisation, both in industry (mainly content publishers, localisation service and tool providers, translators) and in academia (mostly CNGL[12] researchers). The importance of determining what metadata is required was

[11] http://wiki.oasis-open.org/oaxal/, 10/06/10.

[12] CNGL stands for 'Centre for Next Generation Localisation' and is the project the authors are working in.

highlighted by many respondents. Lack of converters and not full support of XLIFF by current localisation tools were some other comments, while simplicity, modularisation, and synergy with other standards are desired needs.

At another, more generic, level, we would like to finish this part of the standards section with the EN 15038:2006 standard. British Standards Institution (BSI) [8] defines it as follows:

> *"EN 15038:2006 specifies the requirements for the translation service provider with regard to human and technical resources, quality and project management, the contractual framework, and service procedures."*[13]

This definition summarises that this standard pertains to human resources (management) and professional competences of translators/revisers/reviewers, and also translation, checking, proofreading, review as such. De Angéli [9] describes his own general appreciation and review of EN 15038:2006, and also compliance benefits to translation service providers (TSPs). He stresses that only very few useful guidelines are provided for the end-user.

Quality should be the main focus of all standards and as De Angéli [9] points out, it is difficult to define what kind of quality:

- quality of final product or translator;
- quality of final revision/check procedure;
- quality of processes for selecting translators and/or subcontracting;
- quality of managing the whole translation process.

This is why there are different standards for different language resources, tools, technologies, stages, and people involved. As seen from the previous paragraphs, there are standards related to authoring, internationalisation, translation, and localisation, but we believe that they are useful under the condition that they are well specified, adopted, and fully supported. Open standards are a good way to make tools and resources interoperable, and facilitate communication between people at different stages of a localisation workflow. XLIFF, being an open standard supporting interoperability, is reviewed in detail in the next section.

3 XML Localisation Interchange File Format (XLIFF)

XLIFF, as its name implies, is an XML-based file format which stores and carries localisation data, and is an interchange, intermediate format. According to XLIFF specification[14]:

[13] http://shop.bsigroup.com/en/ProductDetail/Post.aspx?id=161025&epslanguage=EN&pid=000000000030122045, 10/06/10.

[14] http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html, 10/06/10.

> *"XLIFF has been designed by a group of software providers, localisation service providers (LSPs), and localisation tool providers. It is intended to give any software provider a single interchange file format that can be understood by any localisation provider."*

XLIFF was created in September/October 2000 as a data container (the initiative was called "Data Definition Group"); the original group included Oracle, Sun, IBM, Novell, Berlitz GlobalNET, and focused on the issue of localisation file interchange. XLIFF joined OASIS in December 2001 and the first version (v1.0) was released in 2002. Version v1.1 followed in 2003 and v1.2 in 2008. In May 2009 the XLIFF TC formally entered requirements gathering stage for XLIFF 2.0. Generally speaking, the TC members discuss specifications, define requirements, and also promote the adoption of XLIFF.

A basic minimal XLIFF file (3) follows:

(3) Minimal XLIFF file with one translation unit (TU)

```
<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:documen
     t:1.2 xliff-core-1.2.xsd'>
  <file original="conference.txt" source-language="en-US"
    datatype="plaintext">
     <body>
       <trans-unit id="#1">
             <source xml:lang="en-US">conference</source>
             <target xml:lang="de-DE">Konferenz</target>
       </trans-unit>
     </body>
  </file>
</xliff>
```

On the first line we have the XLIFF declaration and also the `schemaLocation` attribute of the XML schema-instance namespace. In the `file` element we have the name of the file (conference.txt), its source language (English (US)), and its data type (plain text). Then the translation unit element follows with its source and target text in the source language (SL) and target language (TL), respectively.

XLIFF also allows the storage of important data for (software) localisation; an example is the `restype` attribute. Among its values, are `checkbox`, `cursor`, `dialog`, `hscrollbar` (horizontal scrollbar), `icon`, `menubar` (a toolbar containing one or more tope level menus), and `window`. An example[15] of a dialog resource type follows:

[15] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html#datatype, 10/06/10.

(4) XLIFF TU of a dialog

```
<trans-unit id="34" resname="IDD_ABOUT_DLG"
restype="dialog" coord="0;0;235;100" font="MS Sans
Serif;8" style="0x0932239">
  <source>About Dialog</source>
</trans-unit>
```

In example (4), we see metadata about font, style, and coordinates. This is specific to the dialog resource type. When metadata is about the file and the localisation process in general, then it is included in the `header` element. An example of the header element follows:

(5) Process metadata in `header` element

```
<header>
   <phase-group>
      <phase
            phase-name="engineering"
            process-name="sizing"
            contact-name="John Doe"
            contact-email=foo@example.com/>
   </phase-group>
</header>
```

As we can see, metadata such as phase and process name can be provided, while the name and e-mail of the person involved in this task (in this case the engineer) are also given. In `header`, besides contact information, other metadata could include a link to glossaries and other reference material, information about counts, tool name, and comments.

During the different stages of localisation, the XLIFF document might include data from TM, machine translation (MT), and Term-Bases (TB). Segments can be marked as signed-off, needing review, etc. It also supports metadata such as restrictions on string length, coordinates of dialog boxes, as well as specifying the data type for a translation unit.

(6) XLIFF example with alternative translations

```
<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.xsd'>
  <file original=".txt" source-language="en-US"
   datatype="plaintext">
    <body>
```

```
      <trans-unit id="#1">
         <source xml:lang="en-US">conference</source>
         <target xml:lang="de-DE">Konferenz</target>
      <alt-trans tool="Google-translate">
         <source>conference proceedings</source>
         <target xml:lang="de-DE">Tagungsband </target>
         <target xml:lang="es-ES">actas del
         congreso</target>
      </alt-trans>
      </trans-unit>
    </body>
  </file>
</xliff>
```

In the example (6), we do not only have the translation unit as in (3), but also an alternative translation (`alt-trans`), and even in more than one language (German and Spanish). This alternative translation has been generated by an MT system, and the name of the tool is also provided (`Google-translate`). It is also possible to have alternative translations by TM tools and metadata about matching percentages as well.

To sum up, XLIFF is an OASIS standard whose purpose is to carry localisation data and metadata across different people, tools, and localisation stages. It is a bilingual file format with translation units in SL and TL, while it has multilingual support in alternative translations given by TMs or MT systems.

3.1 XLIFF Limitations

In this paragraph we will discuss how although XLIFF is a viable standard, it has some limitations which prevent tools and tool providers from supporting it. First of all, there is a plethora of namespaced elements and attributes, let alone the user-defined ones. More precisely, in the current version 1.2, there are 37 elements, 80 attributes and 269 pre-defined values. In total, there are 386 items. Although these items cover many different aspects of the localisation process and provide a thorough presentation, it is still a very complex vocabulary for people and applications to understand, recognise, and process. Accordingly, this information is difficult to render with computer-assisted translation (CAT) tools and handle in their graphical user interfaces (GUIs).

Another limitation is interoperability. Many CAT and localisation tools use different flavours of XLIFF, i.e. bad data rendering, which makes the cross-tool operations difficult. Errors can follow, such as loss of data and metadata, user-defined elements that are not “understood” and/or misinterpreted by other tools, and in the worst cases, inability to open files or total file corruption. In addition, there are non-standard compliance elements, a lack of conformance clauses, and violations of the XML schema/XSD (XLIFF 1.2) and DTD (XLIFF 1.0) which are reasons for the lack of interoperability (see Morado, [10]).

A third limitation, which is not exclusively related to XLIFF, but rather to all XML-based and open standards, is their customisation. Two extremes exist here:

inflexibility and overflexibility. The former, strict customisation, can restrict innovation, choices, and capabilities, while the latter, extreme customisation, can lead to a limited support of the standard.

Finally, another general limitation is that people are maybe not aware of the existence of standards. Standards should be promoted, so that they are used by publishers/clients, and as a consequence, tool providers. This would mean, in the case of XLIFF, that professional translators would use XLIFF more and more. With the use of a conformance clause (and its acceptance), also at different levels, successful support can be achieved.

In fact, all of the above limitations are linked to each other. Because the standards' vocabulary is often very complex, tool providers customise it to extremes, so that standards lack their rigid structure. If the rigid structure is missing, then tools and formats are not interoperable. Last but not least, the adoption of standards is very important and has to start from the customer, so that the service and tool providers start to implement and support it, and end-users use it.

4 XLIFF Interoperability Experiment

In order to attain data and metadata acquisition and sharing, interoperability of tools and standards is necessary. As stated in the XLIFF 1.2 White Paper [11], interoperability is one of the main aspects of XLIFF:

> *"One of the primary aspects of XLIFF is to allow data to progress from one step to another using different tools – commercial or in-house. Interoperability was one of the key requirements of XLIFF and it continues to be an important aspect of the 1.2 specification."*

In the following sections we test, through our experiment, whether this interoperability aspect holds true by testing several tools handling XLIFF files. We also analyse the manipulation of the standard by different XLIFF converters and the current adoption of the standard in different CAT tools. Our motivation for this experiment is to evaluate our belief that XLIFF has a complex vocabulary which cannot be easily handled and fully supported by CAT tools. We also want to test the actual implementation of the standard by the tools evaluated and generate recommendations for their future development.

In the next table we present some proprietary CAT tools which support XLIFF and distinguish between those that have XLIFF converters and those that have editors. Converters are also called filters and their functionality, in this case, is to transform different file formats into XLIFF. Table 1 shows a list of the CAT tools used for our experiment:

For the purpose of our experiments, we used the "1st International XLIFF Symposium[16]" webpage (.html), released on 3rd May 2010, as our original file. We used an

[16] http://www.localisation.ie/xliff/, 03/05/10.

Table 1. Tools and XLIFF support

CAT tools	Converter/Filter	Editor
Swordfish II[17]	✔	✔
SDL Trados Studio 2009 Freelance[18]	✔ (SDLXLIFF)	✔
Alchemy Catalyst 8.0[19]	✘	✔

html document because the tools we used can convert this format into XLIFF. The stages of the workflow for the experiment are:

1. Conversion into XLIFF;
2. Validation and analysis of the XLIFF file;
3. Manipulation of the file in a specific XLIFF editor (CAT tool);
4. Extraction and analysis of the code produced by the editor;
5. Back-conversion into the original file and analysis of the output.

We followed the same workflow steps with each CAT tool above and combined them all (when possible) to have a sample of the interoperability between all the possible combinations of filters-editors. The process in each case is shown in the following diagram:

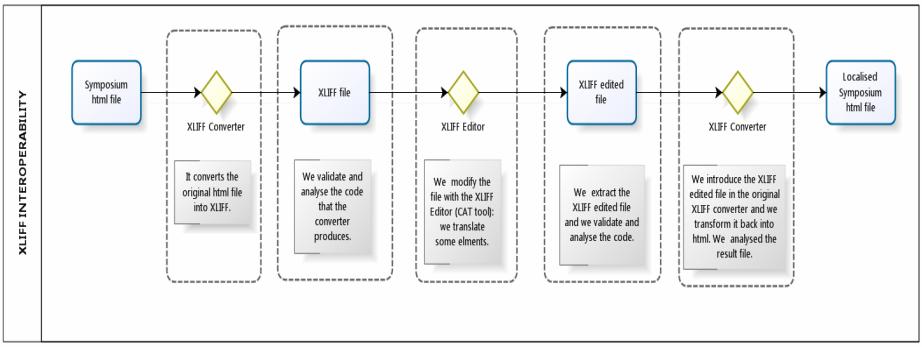

Fig. 2. Experiment workflow

The description of the workflow steps follows here in detail:

1. *Conversion*: We used each one of the XLIFF converters mentioned above to transform our original html into XLIFF.
2. *Validation and analysis:* We extracted the XLIFF file produced by the converter and validated the code using the XLIFFChecker[20] tool. We also analysed

[17] http://www.maxprograms.com/products/swordfish.html, 10/06/10.
[18] http://www.translationzone.com/en/products/sdl-trados-freelance/, 10/06/10.
[19] http://www.alchemysoftware.ie/products/alchemy_catalyst.html, 10/06/10.
[20] http://www.maxprograms.com/products/xliffchecker.html, 10/06/10.

the code in detail to see if there are code features that are worth mentioning. More precisely, a code analysis includes the following steps:

- Version of XLIFF used;
- Metadata contained in header element;
- Information included in translation units;
- Any anomaly in code (which may have been already discovered by the validator).

3. *Manipulation:* We opened another CAT tool (different from the converter used before), which we used as an XLIFF editor to modify our file. We imported the file into the tool and see how it is displayed. We annotated any possible issues in the display of the file. Then, we translated the file into Spanish. We considered the three last paragraphs of the web page (Local Organisation Committee, Local Committee, and Further Information) as "incomplete". As we will see from the next subsections, the tools mark this "incompleteness" in a different way. *Swordfish* uses "needs-review-translation" (in compliance with the XLIFF v1.2 specifications), *SDL Trados* uses "draft" and *Alchemy Catalyst* "needs-review[21]".
4. *Extraction and analysis:* We extracted the modified XLIFF file with all the tools which have this option. We validated the code produced using the XLIFFChecker. Then, we analysed the code produced and we annotated any possible anomalies (we analysed the code in the same way as explained above in *validation and analysis* stage).
5. *Back-conversion:* We sent the modified file back to the original XLIFF converter that we used in the beginning. We saw how the file is displayed and whether the "needs-review-translation" translation units are recognised. Finally we converted the file into its original html format and observed the result.

4.1 SDL Trados Studio (Converter) - Swordfish (Editor)

SDL Trados uses a flavour of XLIFF that differs substantially from the current XLIFF 1.2 version. We validated the file and is a valid XLIFF 1.2 transitional[22] document. The version of XLIFF used is XLIFF 1.2, however a namespace is introduced that refers to its own *SDL Trados* version of XLIFF. It is worth mentioning that the header element contains a lot of metadata: file information, font sizes, colour, and so on. The

[21] "needs-review" is the value that appears in the XLIFF file, but in the *Alchemy Catalyst*'s GUI its equivalent is "for review".

[22] There are two XLIFF "flavours": strict and transitional. Their definitions follow:
Transitional: Applications that produce older versions of XLIFF may still use deprecated items. Deprecated elements and attributes are allowed. Non-XLIFF items are validated only to ensure they are well-formed.
Strict: All deprecated elements and attributes are not allowed. Obsolete items from previous versions of XLIFF are deprecated and should not be used when writing new XLIFF documents. In order for XLIFF documents with extensions to validate, the parser must find the schema for namespaced elements and attributes, and elements and attributes must be valid.
http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html#Intro_Flavors, 10/06/10.

translation units do not have any attribute that denotes their state. The file is imported correctly in *Swordfish* and we marked the last three paragraphs as "needs-review-translation"; the tool has the option to add the state "needs-review-translation" (same as XLIFF's predefined value).

Analysing the file modified by *Swordfish*, the version or the header information has not been modified. The translation units still do not have any state information. Therefore, there is not any difference between the translation units that were marked as approved and those that were marked as "needs-review-translation".

At the back-conversion stage, the file could not be opened by *SDL Trados*. The error message that is produced is that the file does not seem to be a valid XLIFF document.

To sum up, interoperability between *SDL Trados* and *Swordfish* has not been successful, because the former tool cannot open the file that has been modified by the latter.

4.2 Swordfish (Converter) – Alchemy Catalyst (Editor)

The XLIFF file converted by *Swordfish* has been validated and is a valid XLIFF 1.2 strict document. The tool has created 48 translation units (TUs), going from id "0" to "47". All of the TUs have the attribute "approved" with value "no". Importing the XLIFF converted file into *Alchemy Catalyst*, the software localisation tool recognised the file and displayed it correctly. The source text has been automatically copied to the target segments. All the segments appear as "untranslated". *Alchemy Catalyst* has four possible statuses for the strings: "locked", "untranslated", "for review", and "signed off". We translated the file and marked the strings of the first part of the file as "signed off", and the strings of the last three paragraphs "for review"; the option that the tool gives us for this option is "fuzzy"; there is no other state available in the tool.

Analysing the file modified by *Alchemy Catalyst*, it seems to be an invalid XLIFF 1.2 document because the tool introduced the value "needs-review" in the attribute "state". This value was a predefined value of XLIFF v1.0. The value of the attribute "approved" that is present in all the translation units has not been modified (still "no"). However, the segments that were translated and approved by *Alchemy Catalyst* have now the "state" attribute with the value "signed-off". The last three paragraphs, instead of the "signed-off" value, have the "needs-review" attribute. In a nutshell, *Alchemy Catalyst* successfully opened and manipulated the file. However, as it uses XLIFF v1.0 internally, an invalid value for XLIFF v1.2 was introduced ("needs-review") and the approved attribute was not modified.

Converting it back to *Swordfish*, the file opened, but all target segments are marked as "signed-off", and all translation units are marked as "unapproved". The file was converted back to its original html format correctly after a warning that there are unapproved segments.

Generally the interoperability between *Swordfish* and *Alchemy Catalyst* has been relatively successful. There were issues with the attribute "approved" that was introduced by *Swordfish* in all translation units' elements and was not recognised or modified by *Alchemy Catalyst*. *Alchemy Catalyst* uses XLIFF version 1.0 internally and

that means that it introduces items that are not longer valid in XLIFF 1.2 (the mentioned "needs-review" in the attribute "state").

4.3 Swordfish (Converter) – SDL Trados Studio (Editor)

The XLIFF file converted by *Swordfish* is imported to *SDL Trados*; the file is displayed correctly. All segments appear as "not translated". *SDL Trados* offers seven possible statuses for the segments: "not translated", "draft", "translated", "translation rejected", "translation approved", "sign-off rejected", and "signed off". Again we translated the file and marked the strings of the first part of the file as "translated", and the strings of the last three paragraphs as "draft".

Then we validated the file and confirmed that it is a valid XLIFF 1.2 document. The value of the attribute "approved" that is present in all the translation units has not been modified; it is still "no". However, the segments that were translated and marked as "translated" now have the state attribute (inside the target element) with the value "translated". The last three paragraphs, instead of the "translated" value have the "needs-l10n" value (we had marked them as "draft"). To conclude, *SDL Trados* handled and modified the file well without introducing any non-standard compliant items. The only issue is that the value "approved" in the translation unit element has not been modified.

The file was opened correctly in *Swordfish*. The attribute "state" and its values ("translated" and "needs-l10n") have been recognised correctly; as expected all the segments have been marked as unapproved.

We experienced problems with the back-conversion using the filter that *Swordfish* has and the file could not be converted.

In total, the first steps were successful, but full interoperability between *Swordfish* and *SDL Trados* has not been achieved in the last stage of the experiment.

4.4 SDL Trados Studio (Converter) – Alchemy Catalyst (Editor)

The file, after being converted by *SDL Trados*, was imported to *Alchemy Catalyst* and the file was displayed correctly. The TUs that had the attribute "translate" with the value "no" were locked. However, we had to change the file extension for the tool to recognise and treat the XLIFF file correctly. We tried to keep the `<mrk>` inline elements when possible. However if there were already inline elements in the segments we copied the source to target text to keep the logic of the inline elements, so the `<mrk>` elements were deleted.

Alchemy Catalyst has the following statuses: "signed off", "for review", "locked" and "untranslated". After translating, we marked the first segments as "signed off" and the last three paragraphs "for review".

We validated the file and is not a valid XLIFF 1.2 document because *Alchemy Catalyst* introduced the value "needs-review" in the attribute "state" (see 4.3).

The file could not be opened by *SDL Trados* because the validation process failed; it detected the invalid attribute "needs-review" that *Alchemy Catalyst* introduced.

The interoperability between *SDL Trados* and *Alchemy Catalyst* has not been successful. *SDL Trados* cannot open the file modified by *Alchemy Catalyst* because this tool introduced invalid attribute values to it.

5 Conclusion and Future Prospects

In this section we draw some conclusions from our experiment and make future prospects related to a larger scale future experiment. A short summary of the paper and some recommendations about standards follow.

General interoperability results summarised in the various experiment steps (4.1 – 4.4) are depicted in the table below:

Table 2. General experiment results

Converters➔ Editors🡻	**Swordfish**	**SDL Trados**
Swordfish	-	✗
SDL Trados	✗	-
Alchemy Catalyst	✔	✗

The tick ✔ means that the interoperability has been successfully reached without any problems, while the ✗ means that here was an error in some phase of the process and the interoperability has not been reached.

We decided to use a small and simple html document because the tools that we used can convert this format into XLIFF. So many results depend on the specific filter that the CAT Tools have to convert html into XLIFF, and they may be different when dealing with other types of data. Further experiments will try converting different data types. Also, larger and more complex files will be part of our future experiments.

We also plan to interview the developers of the tools used in this experiment. An interview analysing the results of this experiment may answer many of the open questions that are left, may resolve some of the errors, and also may help the future development of the tools.

We did not use any TM or MT system when manipulating the file. A further experiment will require the study of the XLIFF files when using previous material and examining how this is displayed and treated. Also, an analysis of the `<alt-trans>` unit element will be another path to continue with.

In this paper we referred to localisation process and its standards. Starting with the importance of metadata and standardised metadata, we continued to discuss standards for authoring (DITA), internationalisation (ITS), and localisation (XLIFF). The limitations of XLIFF and other standards in general were described. An experiment with an html document, passing it through different CAT tools, covered a practical and important part of the paper. In most cases interoperability was not achieved.

In our opinion, in out-of-ordinary localisation, standards should be open; in the case of XLIFF, more people are involved in projects, more tools are developed, and generally there is independence and wide freedom in selecting the best solution. Standards should be explicit, rich, and also flexible enough, in order to allow tool providers to understand and support them in their tools. Standards should solve a business

problem; having standards and not using them or customising them to extreme is of no help; thus adoption is crucial, although it takes time. Simplicity, better communication between localisation stakeholders, and clarification of specifications are some of our recommendations for current and future standards.

Acknowledgments. 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.

References

1. Schäler, R.: Localization. In: Baker, M., Saldanha, G. (eds.) Routledge Encyclopedia of Translation Studies, 2nd edn., pp. 157–161 (2007)
2. Anastasiou, D., Schäler, R.: Translating Vital Information: Localisation, Internationalisation, and Globalisation. Syn-thèses Journal (2010)
3. Wade, V.: Supporting a Locale of One. In: Proceedings of the LRC Conference XIV, Localisation in the Cloud, vol. 27, p. 27 (2009)
4. ISO/IEC Guide 2:2004, Standardization and related activities – General vocabulary
5. Zydroń, A.: Global Information Management Metrics (GMX), Slaying the Word Count Dragon. The Globalization Insider (2004), http://www.lisa.org/globalizationinsider/2006/04/the_latest_stan.html
6. Cruz-Lara, S., Francopoulo, G., Romary, L., Semmar, N.: MLIF: A Metamodel to Represent and Exchange Multilingual Textual Information. In: Proceedings of 7th International Conference on Language Resources and Evaluation (LREC), Malta, pp. 54–59 (2010)
7. Anastasiou, D.: Survey on the Use of XLIFF in Localisation Industry and Academia. In: Proceedings of Language Resource and Language Technology Standards – State of the Art, Emerging Needs, and Future Developments Workshop, 7th International Conference on Language Resources and Evaluation (LREC), Malta, pp. 50–53 (2010)
8. BS EN 15038:2006, Translation services. Service requirements (2006)
9. De Angéli, G.: Do We Really Need Translation Standards After All? A Comparison of US and European Standards for Translation Services. Translation Journal 12(1) (2008)
10. Morado Vázquez, L., Anastasiou, D., Exton, C.: XLIFF Interoperability Challenges. Poster in: CNGL Scientific Committee Meeting (2010), http://bit.ly/bFt8Zt
11. XLIFF TC. XLIFF 1.2 White Paper (2007), http://www.oasis-open.org/committees/download.php/26817/xliff-core-whitepaper-1.2-cs.pdf