D3.2.3: REPORT ON SHOWCASES

Phil Ritchie (VistaTEC)

Mauricio del Olmo, Laura Guerrero, Pedro L. Díez Orzas, Giuseppe Deriard, Pablo Badía (Linguaserve)

Karl Fritsche, Clemens Weins, Stephan Walter (Cocomore)

Distribution: Public

MultilingualWeb-LT (LT-Web)
Language Technology in the Web

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	3.2.3
Deliverable title:	Report on Showcases
Dissemination level:	PU
Contractual date of delivery:	31st September 2013
Actual date of delivery:	30th October 2013
Author(s):	Phil Ritchie, Mauricio del Olmo, Laura Guerrero, Pedro L. Díez Orzas, Giuseppe Deriard, Pablo Badía, Karl Fritsche, Clemens Weins, Stephan Walter
Participants:	VistaTEC , Cocomore, Linguaserve
Internal Reviewer:	VistaTEC
Workpackage:	WP3
Task Responsible:	VistaTEC
Workpackage Leader:	Cocomore

Revision History

Revision	Date	Author	Organization	Description
1	05/2013	Mauricio del Olmo,	Linguaserve	Initial wording of B2B showcase
2	06/2013	Laura Guerrero	Linguaserve	Manual use of ITS 2.0 metadata
3	06/2013	Pedro L. Díez Orzas	Linguaserve	Business report and revised version
4	06/2013	Karl Fritsche	Cocomore	Completion of CMS sections
8	09/2013	Stephan Walter	Cocomore	Completed "ITS 2.0 annotation and processing in the CMS" and moved it inside "Real experience in localisation with ITS 2.0"
	09/2013	Pedro L. Díez Orzas	Linguaserve	Revised and completed sections 2 and 5.2
		All	All	Revised final version
			Cocomore	Revised and approved final version
			Linguaserve	Revised and approved final version
			VistaTEC	Revised and approved final version
9	10/2013	Phil Ritchie	VistaTEC	Incorporation of Ocelot (Reviewer's Workbench) content.

CONTENTS

1. EXECUTIVE SUMMARY

The present document constitutes a detailed report of the showcases “Improving The Linguistic Review and Post-Editing Process With Ocelot” and “B2B Integration”.

In the “Improving The Linguistic Review and Post-Editing Process With Ocelot” section, VistaTEC addresses how the process of Language Review and Post-editing are significantly improved through the capturing and use of ITS 2.0 Language Quality Issue, MT Confidence and Provenance metadata categories.

In the B2B Integration showcase, Cocomore and Linguaserve will show the applicability of LT-Web ITS 2.0 metadata in a CMS to TMS localization chain.

2. INTRODUCTION

VistaTEC's primary output from this project is a platform independent desktop editor called "Ocelot". The editor can read/write xliff+its files; is based on the Okapi framework and thus integrates seamlessly with the Okapi workflow.

Ocelot implements the Language Quality and Provenance data categories of ITS 2.0. Metadata from these categories that is attached to translation units in the XLIFF file can be rendered alongside those segments in a user definable way through a mechanism of rendering rules. In addition to rendering the metadata, the rules can be used to filter segments according to values of the metadata.

The benefit of a translator or post editor being able to see this metadata within their editing environment is that their task can be directed and made more efficient: hints about errors and information of a segment's provenance can aid the translator in deciding how to edit a segment, if at all.

Ocelot is written in Java which makes it platform independent allowing for maximum deployment.

On the other hand, the contribution to the B2B integration Showcase can be summarized as follows:

- Application of the selected data categories in a real workplace environment within an internal localization workflow and CAT (computer aided translation) tools as part of a client CMS – LSP B2B localization chain.
- Manipulation and usage of the selected data categories in XHTML (XML syntax global rules and HTML syntax local rules) format.
- Real translation and revision of the contents of a web portal (VDMA – Machines for Plastics) from German to French and Chinese.
- Some important benefits are:
    - Tighter workflow integration between LSP and CMS
    - Higher automatic control of the content in the CMS and the TMS (e.g. Translate)
    - Higher manual control in the CMS and the TMS (e.g. Localization Note)
    - More knowledge in the CMS and TMS workflows (e.g. Domain)

All this tasks were made in collaboration, Linguaserve and Cocomore.

Regarding business exploitation, we have to take into account that in spite of CMS-TMS roundtrips is a consolidated technology, it is not standardised. In this way, big size clients can afore defining and implementing custom methods to emulate ITS 2.0, but for SMEs is harder.

Also, clients already implemented CMS-TMS interoperability solutions can benefit from ITS .20, but also new adopters can profit the previous experience concentrated in the ITS 2.0 usage.

Finally , the fact ITS 2.0 works with several standards (HMTL, XML, XLIFF, etc.) can impulse clients to use those and to leverage existing linguistic material and technologies in a easier manner.

3. Improving The Linguistic Review and Post-Editing Process With Ocelot

3.1. Ocelot Features

Ocelot has the following features:

- Reads/writes Language Quality Issue and Provenance ITS 2.0 data categories.
- User configurable rules for defining how metadata is rendered in the sidebar alongside each segment.
- User configurable rules for defining how segments can be filtered according to properties of the metadata.
- User configurable keyboard shortcuts for adding new Language Quality Issue metadata.

3.2. Ocelot Benefits

Ocelot increases the efficiency and accuracy of the linguistic review and post-editing processes by providing translators with access to useful contextual metadata that they may not otherwise have. Metadata gathered at earlier stages of the localization process, such as, automated translation; automated quality assurance checks; NLP-based analytical processes; etc. can inform and direct the translator as to actions that they might want to take.

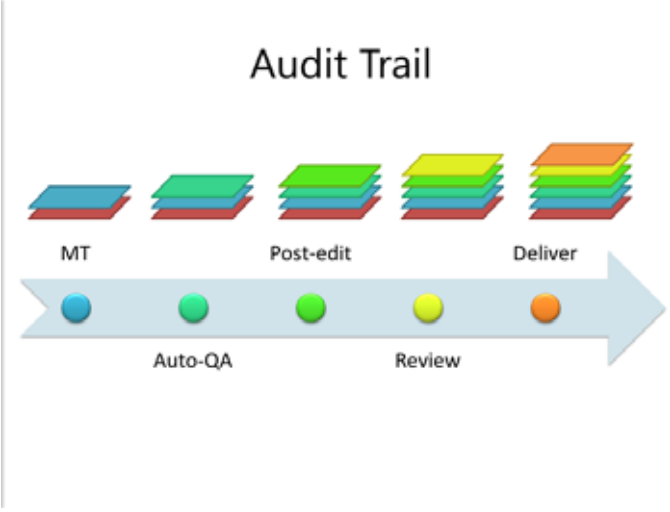
By way of a concrete example:

A document may be translated partially using machine translation and human. At this stage of the process the following Provenance metadata could be gathered:

- Machine translation output confidence scores,
- Name and version of the machine translation engine,
- Name of the human translator and the organization they work for.

Following translation the document could be passed through an automated quality assurance pipeline. During this phase of the localization process the metadata listed below could be added to the segments:

- Name and version of the automated QA programs,
- Types and severity of errors found,
- NLP/Text Classification hypotheses as to the quality and suitability of the translations.



Thus when the document arrives at the desk of a linguistic reviewer, all of this information is at their finger tips. This can enable them to make decisions as to the strategy of their task: address errors first according to severity level, review only those segments proposed by machine translation, specifically review segments proposed by a translator who is known to be a novice, etc.

3.3. Ocelot User Interface

Top left hand panel shows summary of all metadata found in the file. Bottom left hand panel shows segment with raw mark-up. Left hand side of main window shows metadata rendered according to user defined rules.

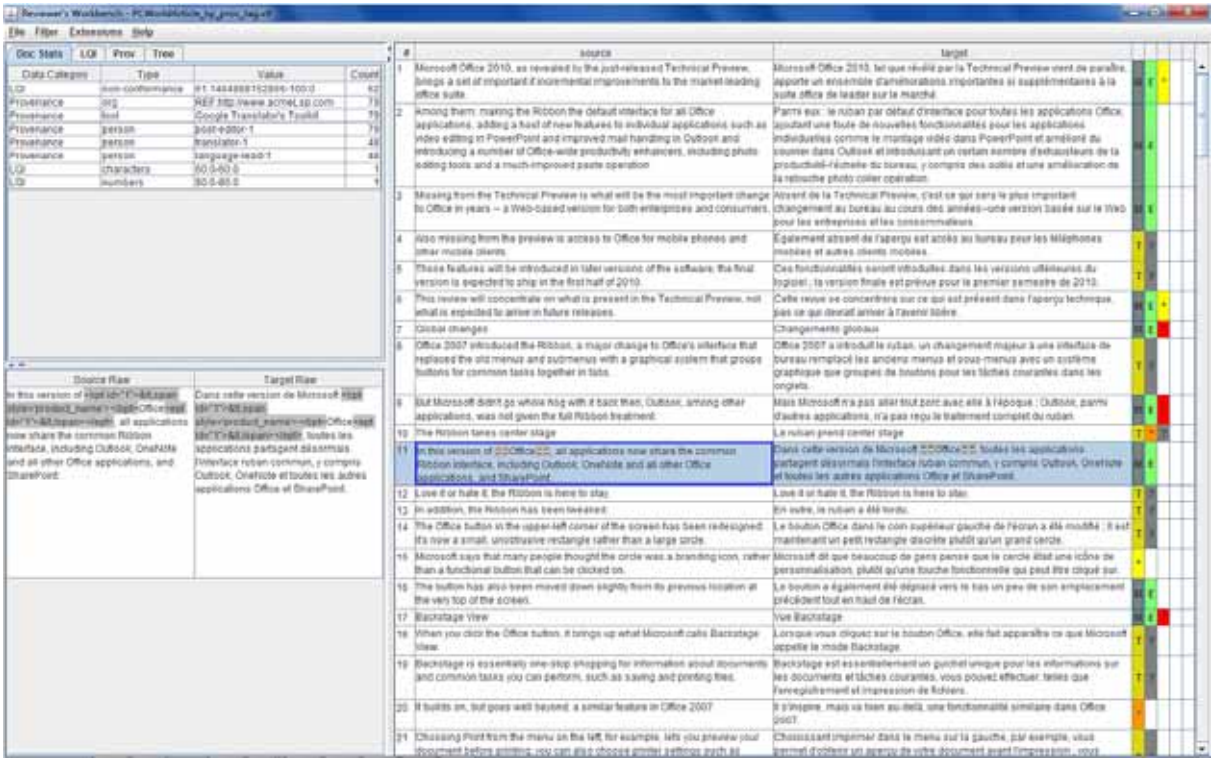
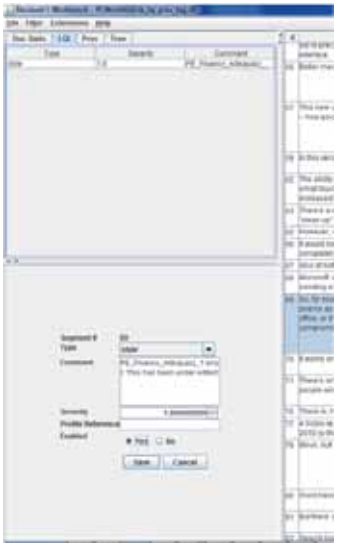
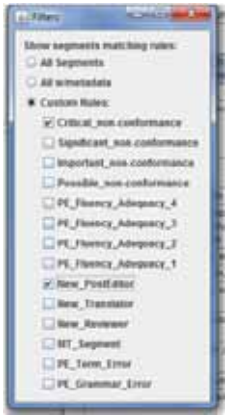


Figure 1: Main Workbench window.



Bottom left hand panel displays the Language Quality Issue data category values associated with the selected segment in the main editor window.



Popup window shows rules associated with particular metadata characteristics used as filtering rules.

LQI metadata can be sent via RESTFul API to VistaTEC endpoint and quality metrics immediately updated in portal dashboard.

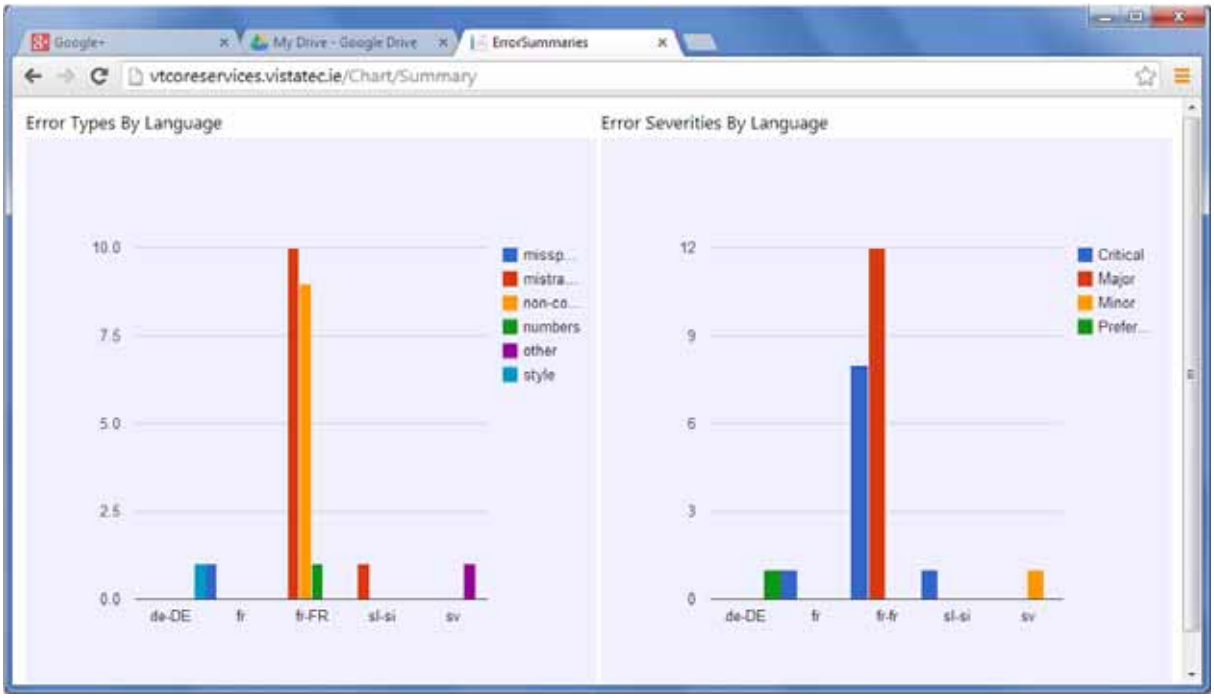


Figure 2: Quality Web Dashboard.

3.4. Making Ocelot Open Source

At Localization World in Santa Clara on 9th October 2013 VistaTEC's announced their contribution of Ocelot as part of the Okapi Framework Project as Open Source under the LGPL License.

3.4.1. RESOURCES

Source Code: https://github.com/vistatec/ocelot

Documentation Wiki: http://open.vistatec.com/ocelot

Binary Downloads: https://code.google.com/p/okapi/downloads/list

4. B2B INTEGRATION SHOWCASE

The large volume of information and web content justifies the use of CMS systems for medium to big size companies and organizations. They provide benefits as content control, several user profiles, abstraction and workflows.

When we introduce the multilingual variable to the CMS picture a translation workflow is highly recommended. The advantages of using an external localization provider and CAT tools gives added value as the use of translation memories, glossaries and the experience with translation management.

This showcase will exemplify how ITS 2.0 allows more integration between the two sides and how the localization workflow of the contents benefits from each data category implemented.

During the project, Linguaserve has provided web service documentation and support to Cocomore for Drupal modules, as well as support in testing.

The contents are generated in Drupal, a Content Management System (CMS). Before they are sent, the contents are annotated with ITS 2.0 metadata in two ways: automatic annotation and manual annotation. XHTML + ITS 2.0 will be used as interchange format. Once created, they are sent to the Linguaserve Global Business Connector Server (GBC Server) translation server, processed in the Linguaserve internal localization workflow Platform for Localization, Interoperability and Normalization of Translation (PLINT). Afterwards, with the annotated content translated and the metadata treated, they are downloaded by the client and imported into the CMS. The ITS 2.0 selected data categories for integration are:

ITS 2.0 Data Category	**Behaviour**
Translate	• *Block parts of untranslatable content*
Localization Note	• *Provide information to translators/revisers* • *Alert the project managers and add tooltip visualization in the workflow*
Domain	• *Provide context to the translators* • *Automatic selection of CAT terminology and translation memories*
Language Information	• *Inform the translators/revisers* • *Update the information after the translation job has been completed* • *Quality check to ensure the source language content complies with the Webservice parameter*
Allowed Characters	• *check if the restrictions are met*
Storage Size	• *Inform the translator/reviser/posteditor* • *Check if the restrictions are met* • *Quality check using the original content*
Provenance	• *Create or update the data category information with the translator/reviser/posteditor who carried out the work*
Readiness (ITS 2.0 extension)* **See section 5.2.1.1*	• *Indicating information about next process to do.*

Table 1: ITS 2.0 Data category applied to CMS-TMS use case

The webservices and interoperability between CMS and TMS recommended several considerations:

a) Webservice definition for Drupal modules.
b) Intercommunication between Cocomore and Linguaserve and testing.

c) Use of data categories first in XML-Drupal format, and finally changed to XHTML-Drupal. They were several interchange format changes (XML -> HTML5 -> XHTML) to cover various needs of the manual task, CMS capabilities, and best practices related to the standard. These changes affected the development of the ITS 2 engine.

The integration on behalf of Linguaserve (Language Services Provider) is being done in three areas:

a) Pre-production/post-production engine for processing content files annotated with ITS 2.0.
b) Linguaserve localization workflow to provide support to project management and production processes.
d) Computer Assisted Translation (CAT) tool usage for translation, revision and postediting with ITS 2.0 annotated content and adaptation of CAT tool filter to ITS 2.0.

Furthermore the real client showcase implementation has required:

a) All data categories implemented are pending final unit tests in coordination with Cocomore.
b) *Domain* workflow integration.
c) *Provenance* workflow integration.
e) Localization case study for ITS 2.0 web localization from German into Chinese and French.
f) Input on best practices in content granularity and analysis of possibilities to provide context to the translators

Other key tasks achieved are:

- Web services connector and engine manipulation unit and integration tests.
- Text annotation. Linguaserve enriches texts (around 75 thousand words) with metadata (support provided by Cocomore).
- Cocomore sends all annotated contents to Linguaserve via web services connector: 75,000 words.
- Translating environment: Linguaserve uses the annotated texts for a human machine-assisted translating scenario from German into Chinese and French (75,000 words per language pair).
- Linguaserve prepares enriched metadata content for Cocomore to import translated annotated texts into Drupal.
- Import of translated and annotated texts back into Drupal begins.
- Import of translated and annotated texts back into Drupal ends.
- Quality assurance: Review and feedback process.
- Linguaserve and Cocomore deliver the first version of a website with annotated and translated text.

Finally, Linguaserve and Cocomore reviewed the whole showcase, this review and web site maintenance tests with ITS 2.0 was required by the customer, including web content and annotation update maintenance undertaken with the full CMS-TMS workflow.

4.1. Real experience in localisation with ITS 2.0

The explanations and conclusions detailed hereinafter are based on the experience of a real localization project with ITS 2.0. Linguaserve, in collaboration with Cocomore, translated more than 60,000 words of the VDMA web page from German into French and Chinese.

First of all, Linguaserve had to train translators and proofreaders in ITS 2.0 in order to understand and take advantage of every tag.

Then we carried out the project from December until February and during this time we had the opportunity to assess the advantages and disadvantages encountered by both project managers and the translators and proofreaders' team.

In the following sections we analyze those tags that have had the greatest impact on the translation and localization process.

4.1.1. ITS 2.0 ANNOTATION AND PROCESSING IN THE CMS

On the content provider's side the ITS 2.0-aware content creation and translation- process defined in the showcase involves the following areas:

- Annotation of source language content with ITS 2.0 metadata within the Drupal CMS.
- Transparent data round-tripping: Triggered from within Drupal, this is realized in the background via export/import of files XHMTL+ITS 2.0 markup, to be sent to/received from LSP. The process is based on an extended version of the Drupal Translation Management (TMGMT)-module.

Additionally, after translation, there's a review step that also happens inside the CMS. ITS 2.0 markup is retained in this step so that annotated information can be taken into account for QA purposes.

We will now focus on the annotation process and provide some details on the different annotation modes and results produced in our test case.

4.1.1.1. ANNOTATION PROCESS

There are two basic types of annotations: Structural annotation rules can be specified as global rules on page/content type level, while local metadata is added by hand. In addition automated annotation tools can be integrated through a standardized interface to support the user in creating such local markup. This feature was however not used in our showcase.

Manual annotation features are available in all generally expected interaction modes (toolbar buttons, context menu, keyboard shortcuts).
Two different manual annotation approaches are supported:

a) Annotation may be done as part of the content creation process, via features that have been added as plugins to the out-of-the-box Drupal WYSIWYG editor.
b) Annotation may be carried out as a separate step, without the ability to modify the content. This allows workflows that separate content know-how and translation management.

ANNOTATION DURING CONTENT CREATION

ITS 2.0 annotation during content creation happens inside the Drupal-WYSIWYG editor, which was provided with new buttons to allow the user to add and edit local ITS markup in content pages. The following ITS data categories can be set with the WYSIWYG while creating or editing a content page:

- Translate
- Locale Filter
- Text Analysis
- Localization Note
- Language Information
- Directionality
- Terminology

Next to the possibility to set these data categories as local markup there are also a few data categories that can act as global markup. Support for such global markup is managed on a per-content-type basis. Enabling ITS support for a given content type creates a new field set in the edit form for content of this type. The field set can be used to enter global XPath rules. It is possible to set default global rules for each content type or globally for the complete site.

For global markup the following data categories are available:

- Domain
- Translate
- Localization Note
- Revision/Translation Agent (from the Provenance data category)x

ANNOTATION AS A SEPARATE WORKFLOW STEP

A second annotation mode is available through a separate tab on the Drupal "Language Management" form. The form provides an editor in it is only possible to work on (add, remove change) the ITS 2.0 markup of a node, while the actual content is all write-protected. This supports a separation of content editing and ITS 2.0 annotation into two distinct workflow steps: A special user role (e.g. a translation manager) can add ITS data very easily after content creation without accidentally changing the content itself. This role will also be able to see and can edit the global markup. The editor provides separate highlighting of local and global markup, annotation through selection and context menus, as well as keyboard shortcuts.

CATEGORIES WITH AUTOMATICALLY DETERMINED VALUES

Some categories allow automatic value determination. This may be because the CMS provides specific means for handling the out of the box, or because adequate values for the can be derived automatically from other information that is available from various sources within the CMS and workflow. This fact is exploited wherever possible. For instance data categories like *Allowed Characters*, *Storage Size* and *Readiness* from the ITS extension will be added automatically to the content sent to the LSP depending on Drupal's field definitions of a particular field. As an example there is a maximum length of 255 characters for the title field, and in this case the *storage size* category is added to the title field with the respective values set. As another example the *expected finalization date* and *priority* are added by the translation manager before the translation job is submitted to the LSP.

4.1.1.2. STATISTICS FROM TEST CASE

In this section we present some figures on the data used in our test case and on the annotations performed in the source text within the Drupal CMS. Table 2 shows some general characteristics of the source text. With some 14 tokens/sentence the average sentence length is rather on the low side for German. This is not surprising however since the texts are press releases and therefore authored for readability.

Total Documents	*141*
Total Sentences	*5228*
Total Tokens	*75496*
Avg. Sentences/Document	*37*
Avg. Tokens/Sentence	*14*

Table 2: Document Statistics for Test Case

Table 3 shows some general statistics on the intensity of tagging in the source texts. With an average of about one tag per sentence (and in fact even almost one manually added tag per sentence), one can say that the possibility of adding meta-information to the sentence has been quite heavily used.

Total Tags	*5544*
Manual	*4700*
Avg. Tags/Document	*39,3*
Avg. Tags/Sentence	*1,0*

Table 3: Tag Usage – Overview

Table 4 shows details on the tag usage for the single data categories. It shows that by far the largest proportion are indications that certain text spans should not be translated (proper names, specific terminology, addresses). The second prominent category are free-text notes to the translator, occurring almost twice per document and covering a total of about 12% of the tokens in the input text. The average note length of about 8 words may be taken as a rough indication that many of the notes were of non-trivial content.

Category	Count	Avg. Count/ Doc.	Avg. tokens annotated	Avg tokens/ note
Translate (value: no)	*4346*	*30,8*	*2,8*	
Localization Note (type: descrip	*238*	*1,7*	*24,9*	*8,1*
Language	*116*	*0,8*	*1,7*	
Allowed Characters	*703*	*5,0*	*104,4*	
Storage Size	*141*	*1*		

Table 4: Tag Usage – Details

All in all we take the observed usage pattern as a confirmation of a central assumption made in the definition of the ITS 2.0 standard: That it makes sense (and actually serves an urgent need) to enable communication information pertaining to the translator's individual translation choices right in the place where it will be applied, i.e. together with the actual text spans that it pertains to.

During localization workflow, Domain (automatically annotated in Drupal) is processed in the TMS. The Translation Management System generates Provenance and Readiness data category. The numbers shown below are obtained from the processing of the content files in both languages (French and Chinese). The number of total processed contents is 284, 142 per language.

Data category	Number of processed tags	Per language
Translate	*8.694*	*4.347*
Localization note	*476*	*238*
Domain	*284*	*142*
Language information	*544*	*272*
Allowed characters	*1.412*	*706*
Storage size	*284*	*142*
Provenance	*568*	*284*
Readiness (*)	*284*	*142*
TOTAL:	*12.546*	*6.273*

Table 5: Tag Usage in localization workflow – Overview

The average of processed tags per file was:

Data category	Average of processed tags per file
Translate	*30,61*
Localization note	*1,68*
Domain	*1*
Language information	*1,9*
Allowed characters	*4,97*
Storage size	*1*
Provenance	*2*
Readiness (*)	*1*
TOTAL:	*44,17*

Table 6: Tag Usage in localization workflow – Overview

4.2. Translator's Training

The usage and implications of the applied ITS 2.0 data categories for the translators and proofreaders are shown in the following screenshot of a file imported in the CAT tool:

ITS 2.0 guidelines Legend

1) Domain: it is used to identify the topic or subject of a given content. In this case it is used to provide context for the translation work. It could provide several values. The text is blocked and readable (Allgemein, Angewandte Wissenschaft, Anlagenbau, Kunststoff- und Gummimaschinen…)

2) Localization Note (global use): it is used to give information to localizers. It provides valuable information and context for the translation work. The text is blocked and readable (Pressemitteilung). In this case the note applies to the whole content.

3) Storage Size: it is used to specify the maximum storage size of a given content. The information is provided in the xMax attribute (xMax="255") and the size of the original text in the xMaxOrig attribute (xMaxOrig="48"). The translation should respect the maximum storage size.

4) Translate: it gives information on whether the content of an element should be translated or not. This could apply only to a particular part of a sentence. In this cases, the parts of a text marked as not translatable will be blocked by Transit, as it is shown in the screenshot where the names of companies, associations and people are marked as not translatable and therefore blocked (VDMA, Marc Wiesner).

5) Localization Note (local use): it is used to give localizers information on a particular portion of content. It provides valuable information and context for the translation work. The text is blocked and readable (Das Internet macht vieles transparenter…) In this case, the note applies only to a particular part of the text.

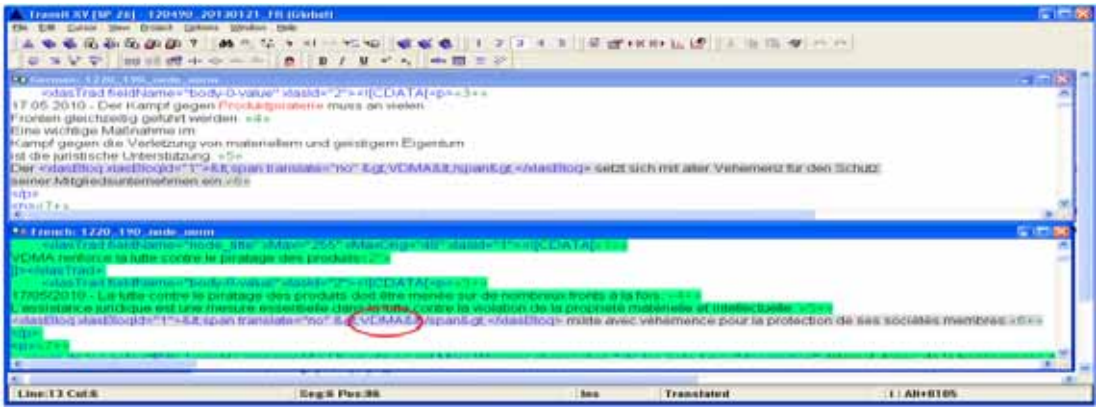
6) Language Information: it is used to indicate the language of a given piece of content. Some parts of the content can be in another language. This is indicated by the lang atribute (lang="en" -> english, Know-how).

ITS 2.0 guidelines sample

4.3. ITS 2.0 translation usage evaluation

4.3.1. TRANSLATE

See http://www.w3.org/TR/its20/#trans-datacat.This metadata specifies whether a piece of content should be translated or not, such us proper names, for instance. Our normalization engine blocks these pieces of content so translators cannot edit them, as it is shown in the following screenshot, where text in black is translatable and text in blue are tags.



In the sentence “Der VDMA setzt sich mit aller Vehemenz für den Schutz seiner Mitgliedsunternehmen ein”, the term VDMA has been annotated as not translatable so our normalization engine has blocked it not to be editable by the translators.

4.3.1.1. ADVANTAGES

1) Translators can have a quick and clear idea of the client’s criteria about which terms they do not have to translate.

2) It promotes terminological consistency, what becomes especially relevant in projects of great volume in which several translators are involved.

4.3.1.2. DISADVANTAGES ENCOUNTERED

1) Too many tags in the same sentence reduce the proper visibility of translatable text.

Example: Text without “translate=no” tags

„Die Attraktivität der deutschen Produkte hat ihre Kehrseite", berichtet Thorsten Kühmann, Geschäftsführer des VDMA Kunststoff- und Gummimaschinen, aus der jüngsten repräsentativen Umfrage des VDMA zur Produkt- und Markenpiraterie…

Example: Text with “translate=no” tags

„Die Attraktivität der deutschen Produkte hat ihre Kehrseite", berichtet <xlasBloq xlasBloqId="7">Thorsten Kühmann</xlasBloq>, Geschäftsführer des <xlasBloq xlasBloqId="8">VDMA</xlasBloq> Kunststoff- und Gummimaschinen, aus der jüngsten repräsentativen Umfrage des <xlasBloq xlasBloqId="9">VDMA</xlasBloq> zur Produkt- und Markenpiraterie…

2) The change of order of a sentence when it is translated into another language usually implies changing the order of tags as well. Considering that these tags are or can be blocked in the CATA tool, this forces translators to unblock such tags and move them while translating, what considerably slows the translation process.

Example: Original text in German

Energieeffizienz und höchste Anforderungen an die Oberflächengüte stehen während der <xlasBloq xlasBloqId="31">Fakuma</xlasBloq> im Fokus von <xlasBloq xlasBloqId="32">ENGEL automotive</xlasBloq>.

Example: Translation into French

Le rendement énergétique et un traitement de surface parfait au centre des attentions pour <xlasBloq xlasBloqId="31">ENGEL automobile</xlasBloq> à l'occasion du salon <xlasBloq xlasBloqId="32">Fakuma</xlasBloq>.

Example: Translation into Chinese

在 <xlasBloq xlasBloqId="31">Fakuma</xlasBloq> 展会上，<xlasBloq xlasBloqId="32">ENGEL automotive</xlasBloq> 将重点放在能源效率和超高的成品质量要求上。

3) Translators must be very attentive to ortotypography, especially when there are several tags in a sentence, in order to avoid lack of spaces, double or wrong spaces…

Example:

En coopération avec son client <xlasBloq xlasBloqId="33">Gerhardi</xlasBloq> (<xlasBloq xlasBloqId="34">Ibbenbüren</xlasBloq> / Allemagne), <xlasBloq xlasBloqId="35">ENGEL</xlasBloq> produit des anneaux décoratifs pour les grilles de calandre de <xlasBloq xlasBloqId="36">BMW</xlasBloq> à l'aide d'une <xlasBloq xlasBloqId="37">ENGEL duo 2550/500 pico</xlasBloq>. <xlasBloq xlasBloqId="38">Gerhardi</xlasBloq> est un spécialiste des ornements ultra-brillants dans l'habitacle ou à l'extérieur des véhicules et dirige un des plus grands sites de galvanisation en Europe.

Example: Translation into Chinese

在与客户 <xlasBloq xlasBloqId="33">Gerhardi</xlasBloq>（总部位于德国 <xlasBloq xlasBloqId="34">Ibbenbüren</xlasBloq>（伊本比伦））的合作中，<xlasBloq xlasBloqId="35">ENGEL</xlasBloq> 将采用 <xlasBloq xlasBloqId="36">ENGEL duo 2550/500 pico</xlasBloq> 为 <xlasBloq xlasBloqId="37">BMW</xlasBloq> 散热器格栅生产装饰环。

In this text, translators have to keep in mind that "Ibbenbüren" is part of the sentence to be translated in order to leave the proper space between the closing tag and the slash.

They also have to check if there is any punctuation sign between tags, such as the point between tags 37 and 38, so as to leave the correct spaces before or after the sign, especially when they have to move the order of tags.

4) Translators must also be very attentive to the concordance of not translatable texts.

Example:

> Le pack de productivité pour les machines polyvalentes <xlasBloq xlasBloqId="31">Golden Edition</xlasBloq> se compose de l'<xlasBloq xlasBloqId="32">Arburg Energiesparsystem (AES)</xlasBloq> avec entraînement de pompe à vitesse de rotation variable et moteur d'entraînement refroidi à l'eau.

This example shows the necessity of using "l'" in this case instead of "le" since the next word, which has been blocked because it has not to be translated, begins by a vowel. Because of the reduced visibility of theses blocked terms it would be quite feasible for the translator to use "le".

5) Wrong application of this metadata can cause serious translation mistakes. One of the terms that were decided to be annotated with this metadata were trademarks, which are supposed to be identical in any language. Nevertheless, in this case the original language was German and the target language was French. In German nouns, even proper nouns such as trademarks, can be declined so despite the trademark remains the same, the word can be altered by any kind of suffix and cannot be left like that in French.

Example: Original text in German

> „Damit ist nach zwei sehr guten Export-Jahren der Einbruch aus 2009 ausgeglichen und der Rekordwert von 2008 deutlich eingestellt", erklärt <xlasBloq xlasBloqId="5">Thorsten Kühmann</xlasBloq>, Geschäftsführer des <xlasBloq xlasBloqId="6">VDMA-Fachverbandes</xlasBloq>.

Example: Translation into French and Chinese

> « Ainsi, après deux très bonnes années au niveau des exportations, le recul de 2009 est compensé et la valeur record de 2008 clairement atteinte », déclare <xlasBloq xlasBloqId="5">Thorsten Kühmann</xlasBloq> le gérant de <xlasBloq xlasBloqId="6">VDMA-Fachverbandes</xlasBloq>.

> "连续两年的高出口额弥补了 2009 年经济危机带来的影响，并超越了 2008 年的记录"，<xlasBloq xlasBloqId="5">VDMA-Fachverbandes</xlasBloq>（德国机械设备制造商联合会）塑料和橡胶机械行业协会负责人 <xlasBloq xlasBloqId="6">Thorsten Kühmann</xlasBloq>（托尔斯滕· 屈曼）说。

This text would be viewed in the web like this:

> « Ainsi, après deux très bonnes années au niveau des exportations, le recul de 2009 est compensé et la valeur record de 2008 clairement atteinte », déclare le gérant de VDMA-Fachverbandes.

“连续两年的高出口额弥补了 2009 年经济危机带来的影响，并超越了 2008 年的记录”，VDMA-Fachverbandes（德国机械设备制造商联合会）塑料和橡胶机械行业协会负责人 Thorsten Kühmann（托尔斯滕· 屈曼）说。

While the correct translation into French would be:

« Ainsi, après deux très bonnes années au niveau des exportations, le recul de 2009 est compensé et la valeur record de 2008 clairement atteinte », déclare le gérant de VDMA-Fachverban.

“连续两年的高出口额弥补了 2009 年经济危机带来的影响，并超越了 2008 年的记录”，VDMA-Fachverband（德国机械设备制造商联合会）塑料和橡胶机械行业协会负责人 Thorsten Kühmann（托尔斯滕· 屈曼）说。

If the language pair would have been Spanish into French, for example, where a trademark would be invariable in both languages, the application of this metadata to trademarks would have worked correctly.

4.3.1.3. CONCLUSIONS

The use of this metadata certainly has great benefits, such as the translation consistency assurance, although the disadvantages encountered reveal the necessity of considering the following recommendations in order to obtain the best results from its application:

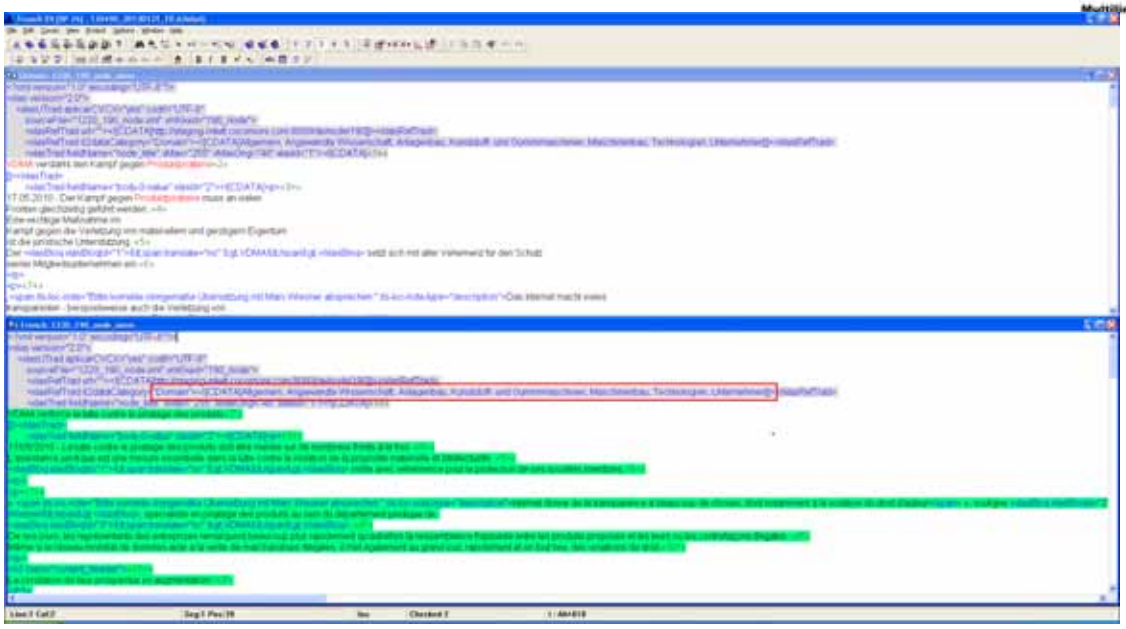
1) Not to abuse of these tags to avoid reducing the visibility of the text to be translated.

2) Try to apply them to isolated terms instead of long sentences to avoid possible ortotypography mistakes.

3) Agree the application criteria with linguistics considering the terminological and grammatical rules of both original and target languages.

4.3.2. DOMAIN

See http://www.w3.org/TR/its20/#domain. This metadata is used to identify the domain of content. The following screenshot shows how this information is viewed in the CAT tool.



4.3.2.1. ADVANTAGES

1) Translators can see the context and thematic area of a particular content.

2) Project managers can also see this information which allows them to:

   a) Group files in different batches for translation according to their domain.

   b) Select the appropriate glossaries and reference materials for each batch of files.

   c) Select the appropriate translators according to their specialization.

3) Synchronization of the domains and the thematic areas with which the terms of the glossaries created for the project have been categorized.

For this project, ten possible domains were defined:

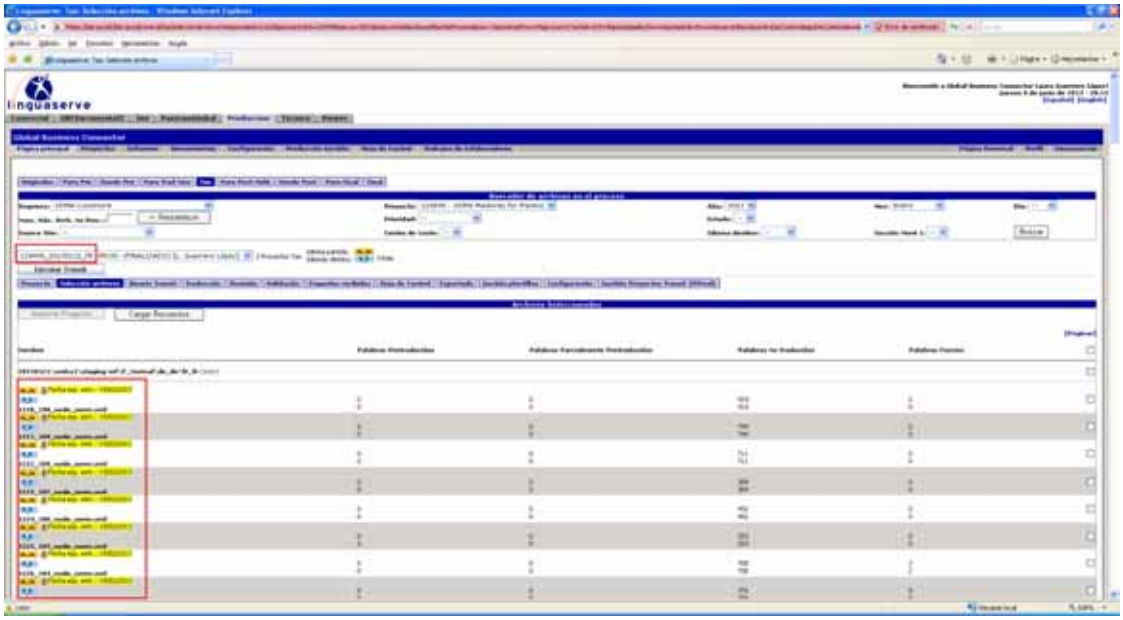
- Allgemein
- Angewandte Wissenschaft
- Sozialwissenschaften
- Technologien
- Unternehmen
- Kunststoffindustrie
- Maschinenbau
- Anlagenbau
- Kunststoff- und Gummimaschinen
- Wirtschaftswissenschaften

And five glossaries were created according to the thematic area of the terms they contain:

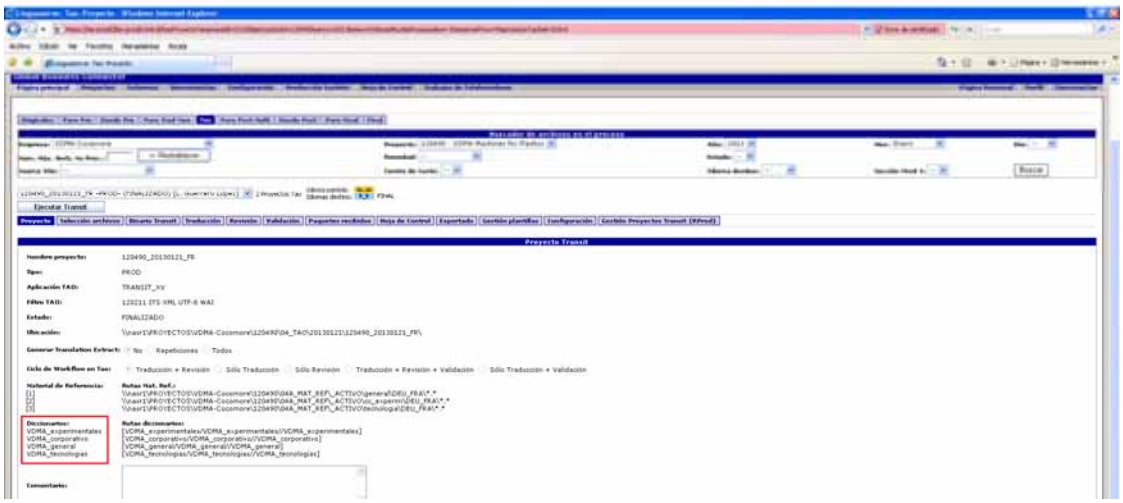
- General (general)
- Ciencias experimentales (experimental sciences)
- Ciencias sociales (social sciences)

- Tecnología (technology)
- Corporativo (corporative)

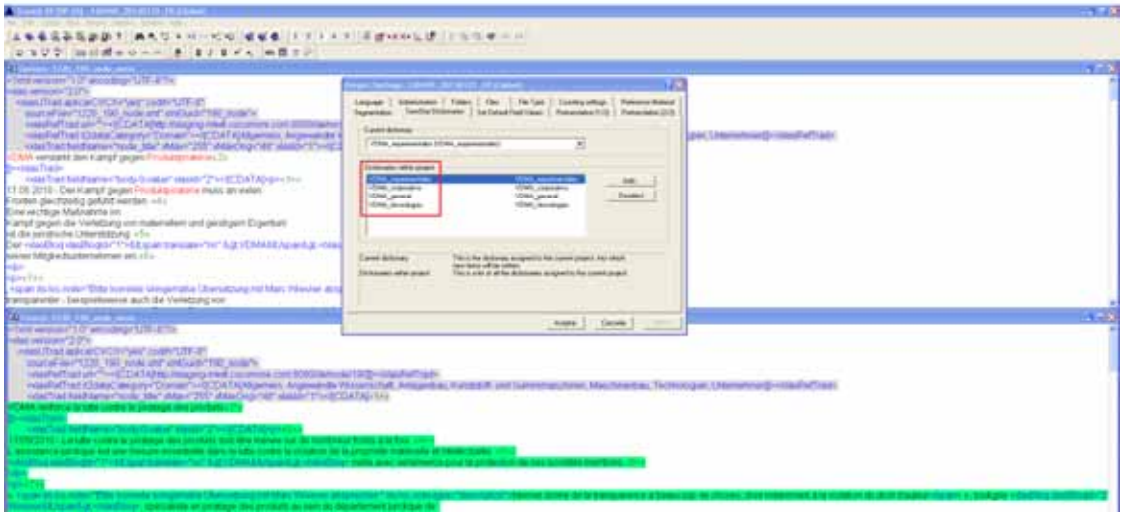
By matching both the domains and the thematic areas, Linguaserve's localization and internationalization platform identifies the domain of each file and selects the appropriate glossary when project manager creates a CAT project containing a specific group of files. The following screenshots show this process:



The CAT project (120490_20130121_FR) contains a group of files to be translated into French.



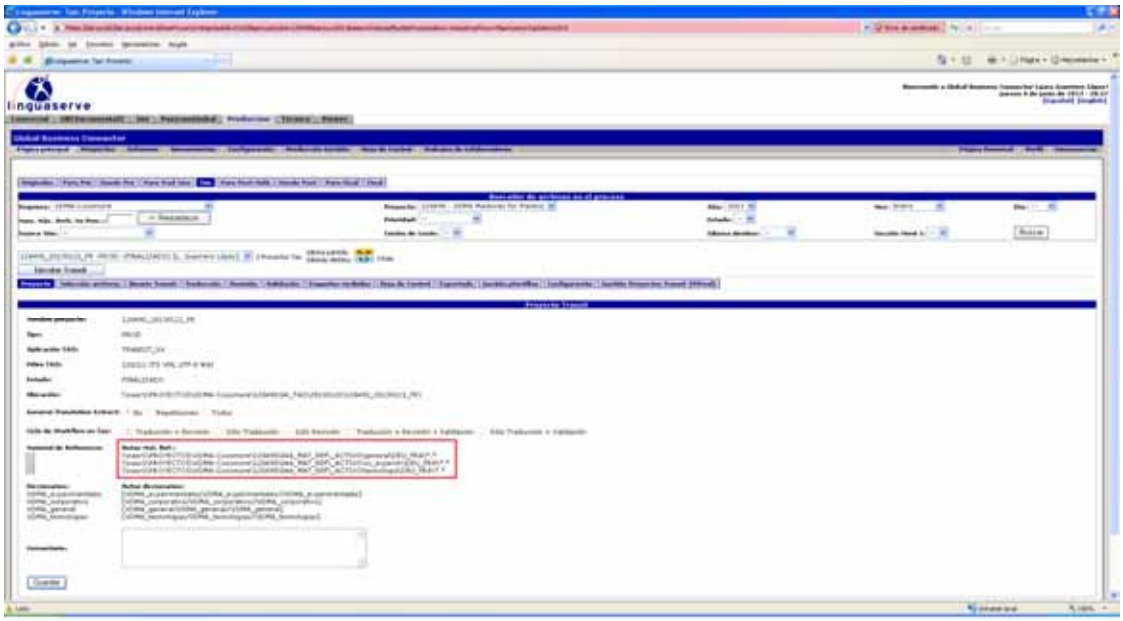
The platform identifies the domain of that batch of files and selects the appropriate glossaries to be used for that particular CAT project.



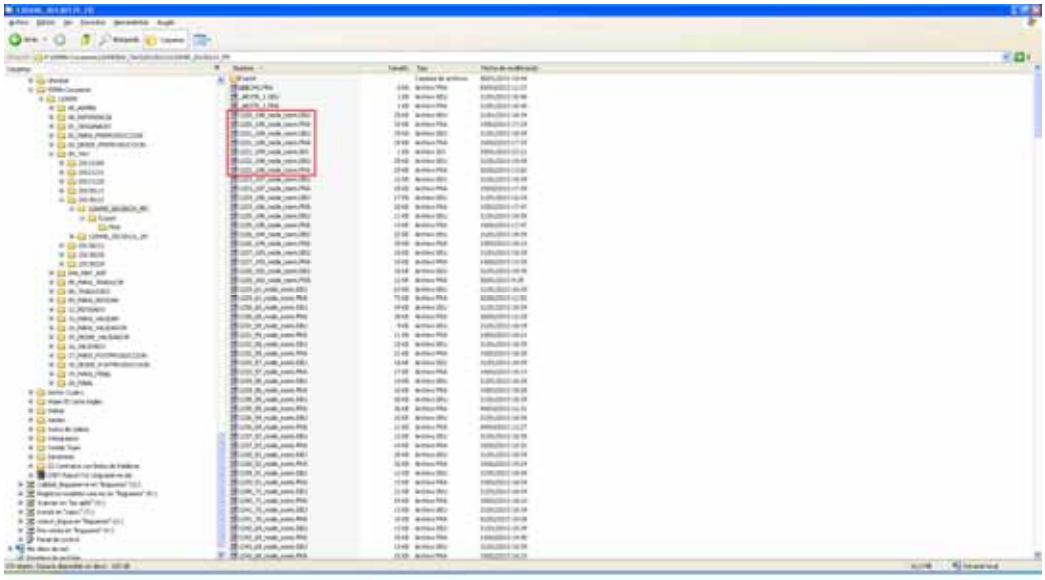
When translators open the CAT project, they receive these glossaries.

4) Synchronization of the domains and the folders where pairs of files (original and translated) are stored conforming the translation memories.

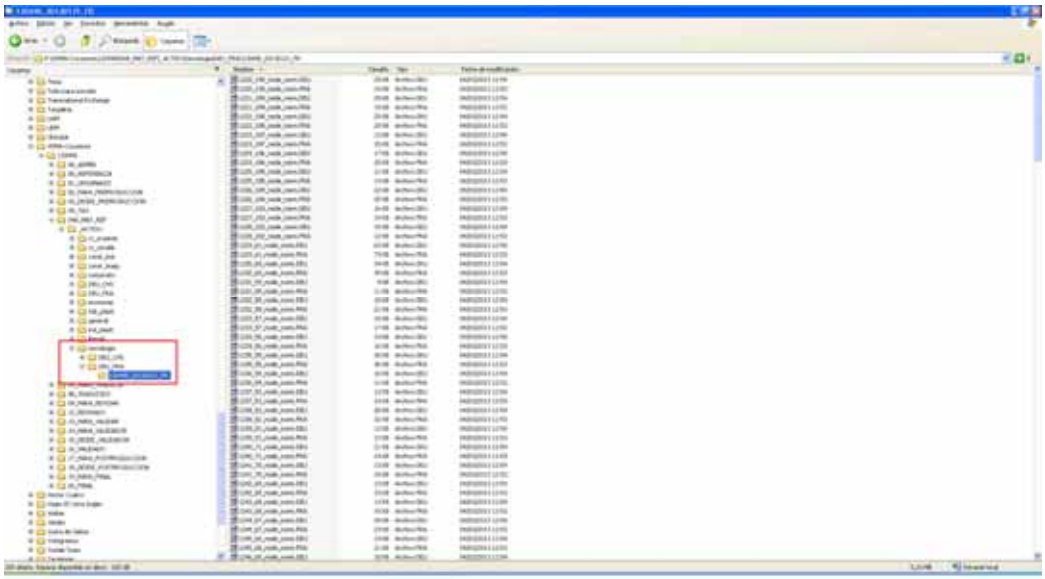
Once a CAT project is finished, i.e. the files it contains have already been translated, revised and exported, Linguaserve's localization and internationalization platform identifies the domain of each file and the pair of files generated by the CAT tool for that particular file are stored in the corresponding folder according the their domain. The following screenshots show this process:



The platform identifies the domain of the files contained in the project and shows the folders where those files will be stored once translated.



The CAT tool generates a pair of files (German and French) for each file.



There is a folder for each of the domains defined and those files are automatically stored in their corresponding folder according to their domain.

This process allows an easy and automatic creation of thematic translation memories that will be selected by the platform when new files for translation are received according to their domain, despite project managers can also add to the CAT project any other translation memory manually for specific purposes.

4.3.2.2. DISADVANTAGES ENCOUNTERED

1) Sometimes a particular file can be classified in various thematic areas so the metadata can contain more than one domain.

If several domains are assigned to a given content, the CAT project created for that file will contain all the

translation memories and glossaries corresponding to those domains so that the translator will lose the benefit of having the specific memory and glossary for that particular project.

In addition, a copy of those files will be stored in all the folders corresponding to the domains assigned to that content, producing the lack of specialization of translation memories which are supposed to be categorized by thematic areas.

4.3.2.3. CONCLUSIONS

The use of this metadata has a great impact on translation quality assurance by offering translators more context and proper reference materials and permitting project managers an adequate and efficient management of translation memories, glossaries and translators' teams.

In order to obtain the best results from its application, the following recommendations should be considered:

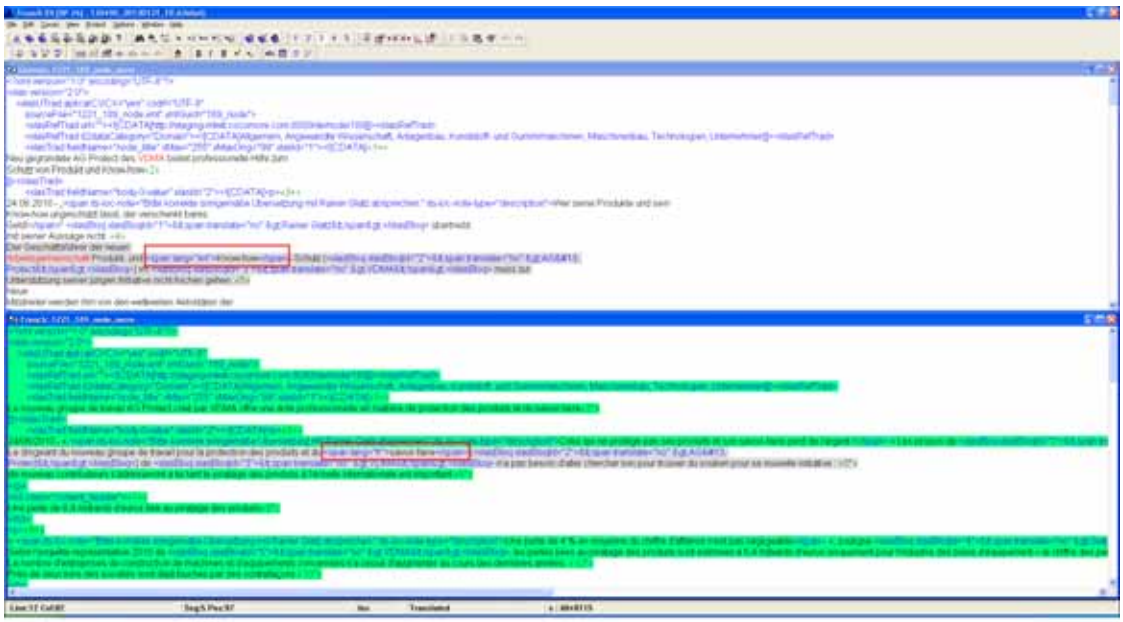
1) Define the domains in collaboration with linguistics (the translation company in this case) once they analyse the contents and define the different thematic areas in which the terms of the glossaries will be classified.

2) Synchronize the domains and the thematic areas defined if different.

3) Try to assign only the domain/s corresponding to the most frequent terminology of a particular content.

4.3.3. LANGUAGE INFORMATION

See http://www.w3.org/TR/its20/#language-information. This metadata is used to indicate the language of a term. Our normalization engine blocks the metadata indicating the language and leaves the term editable to be translated if needed, as it is shown in the following screenshot, where text in black is translatable and text in blue are tags.



In the text in German this tag has been used to inform that “know-how” is an English term.

4.3.3.1. ADVANTAGES

1) It provides information to translators on the language of certain terms used in the text.

2) It allows the application of linguistic criteria in a consistent and reliable manner. For example, in this case it could be used to apply italics to all texts in a language different from French.

4.3.3.2. DISADVANTAGES ENCOUNTERED

1) Translators are forced to unblock the metadata to change the language when the term must be translated into the target language.

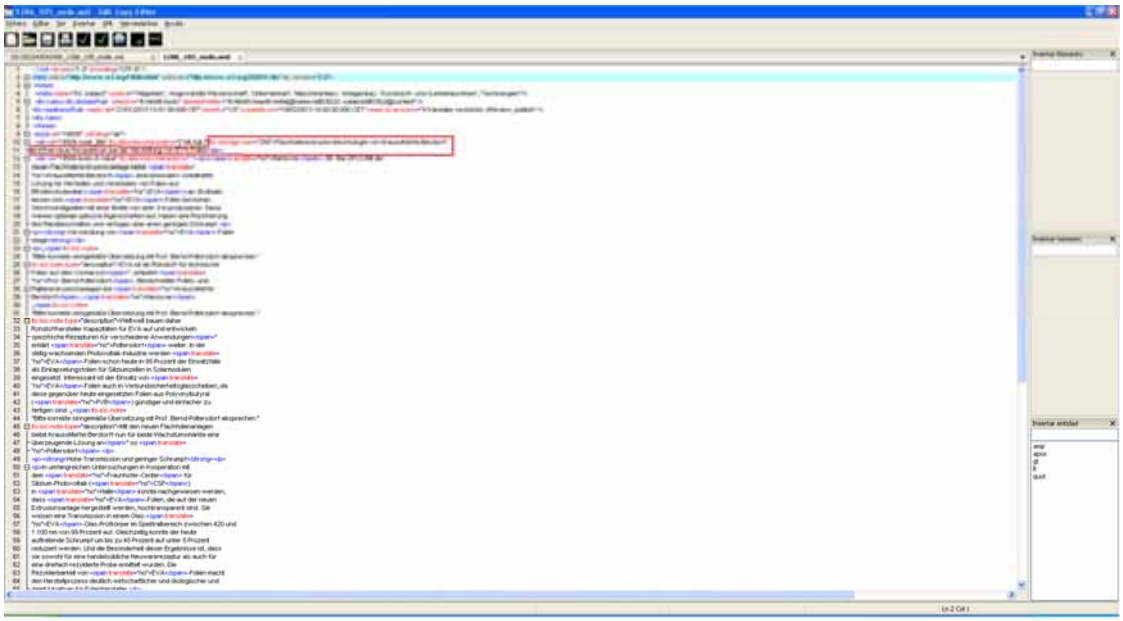
4.3.3.3. CONCLUSIONS

Providing information and context to translators is always positive so the use of this metadata is positive too.

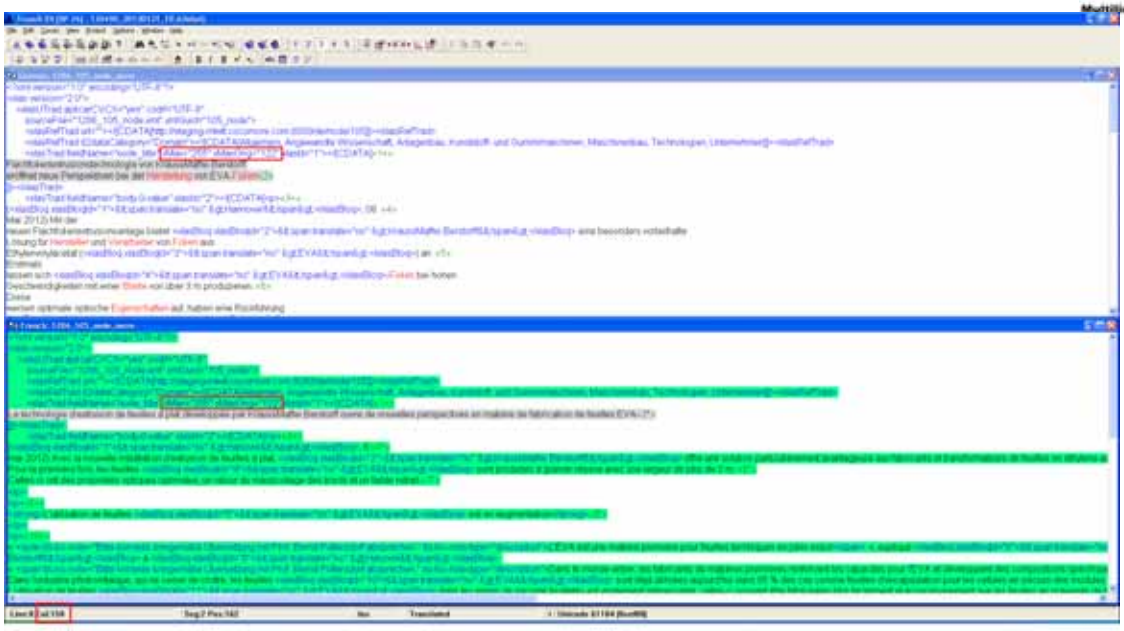
4.3.4. STORAGE SIZE

See http://www.w3.org/TR/its20/#storagesize. This metadata is used to specify the maximum storage size of a given content.

In this project, it normally specifies the maximum size of the title, as it is shown in the following screenshot.



When the file is already processed, this information is shown as follows:



Translators can see the maximum size permitted (255 characters) as well as the size of the original text (122 characters). In addition, they can also see the number of characters of the segment they are translating in the CAT tool status bar (158 characters) so that they can adjust the translation to the maximum size indicated by this tag.

4.3.4.1. ADVANTAGES

1) Translators know the maximum size permitted for the translation of a particular piece of content, what becomes especially relevant in the case of literals. Having this information in advance allows translators adjust the translation in real time, avoiding losing time doing later corrections.

4.3.4.2. DISADVANTAGES ENCOUNTERED

None.

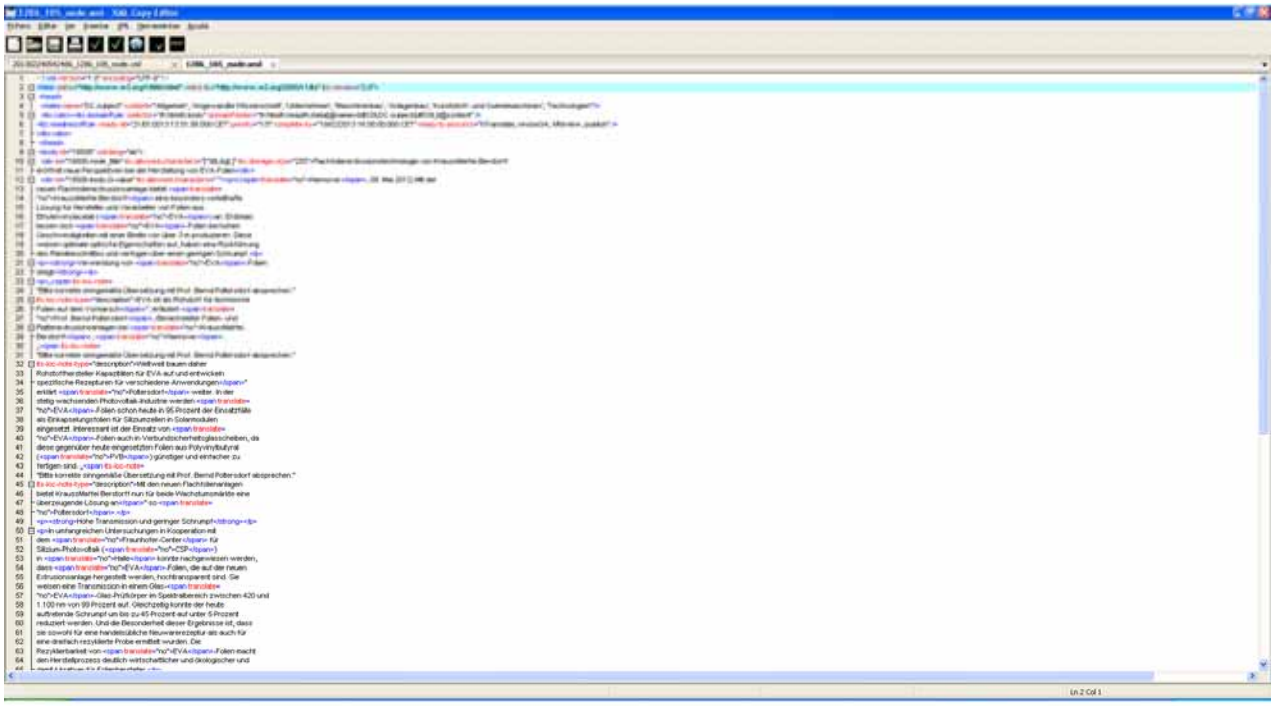
4.3.4.3. CONCLUSIONS

This metadata seems undoubtedly useful, improving and facilitating translation management, especially of those contents for which maximum size is particularly important, such as literals and so on.

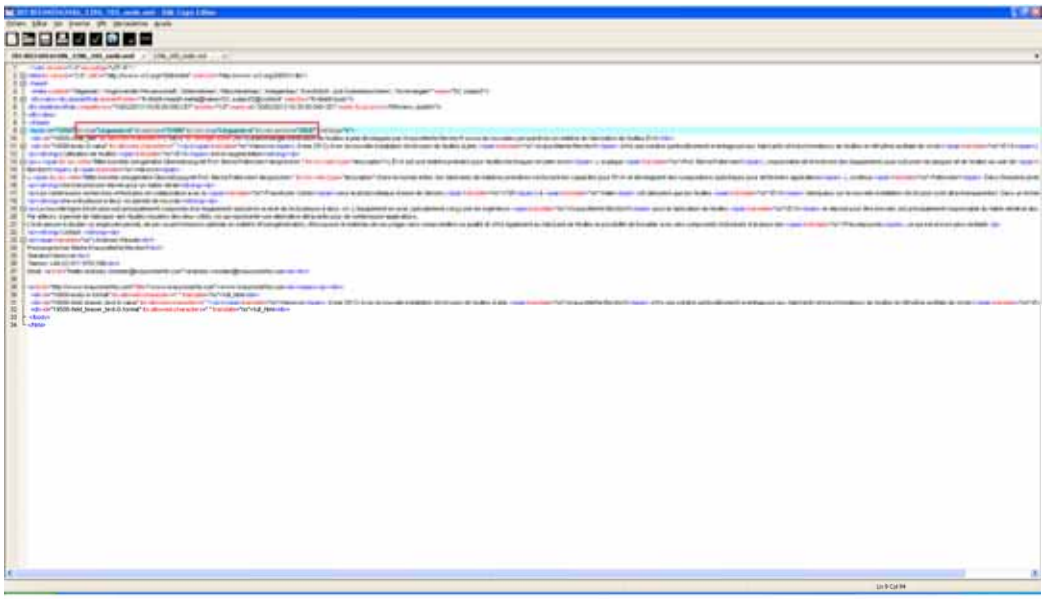
4.3.5. PROVENANCE

See http://www.w3.org/TR/its20/#provenance . This metadata tells who was the translator and the proofreader of a particular file. This information is added to the file once it is translated and revised. The following screenshots show this process:

Original file in German:



Translated file into French:

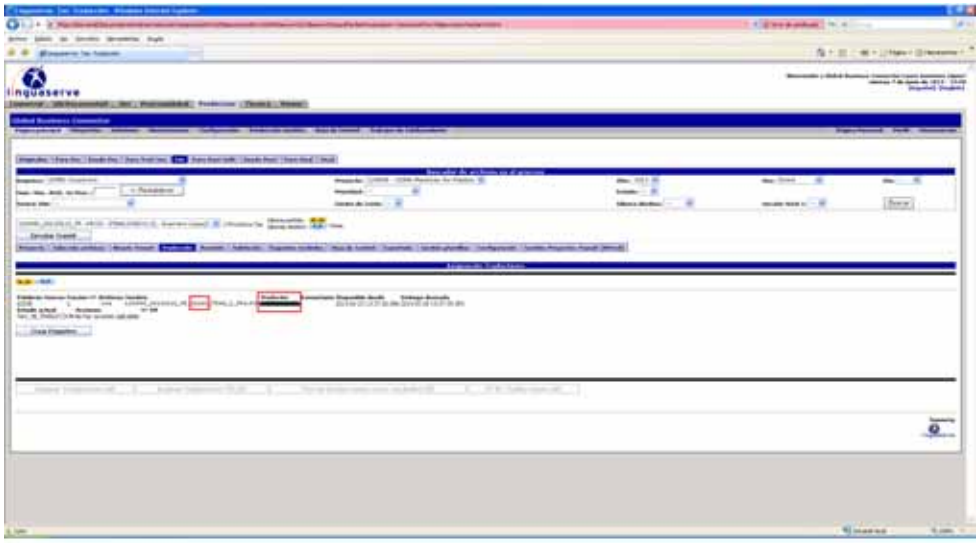


This screenshot shows the following information:

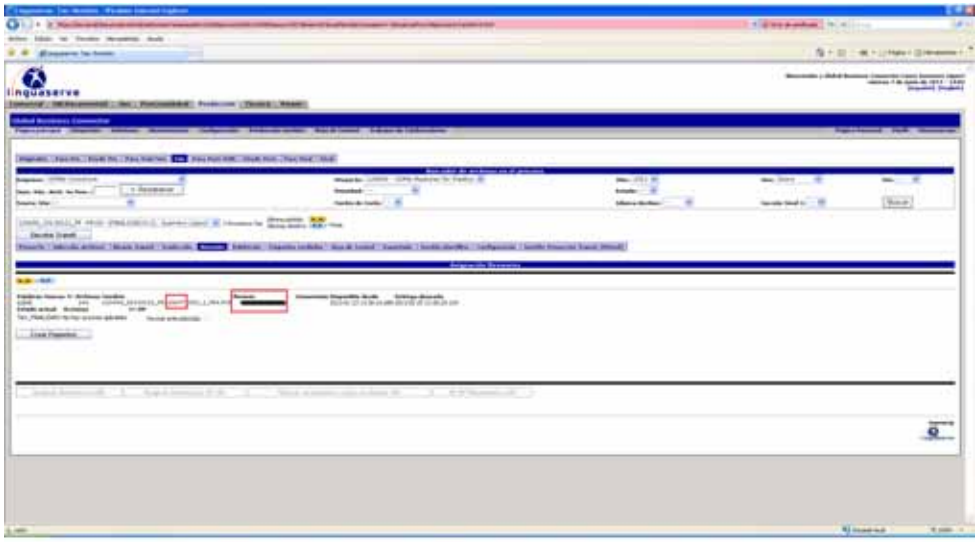
- Its org: company that has carried out the translation (Linguaserve).
- Its person: 21686, being this number the code Linguaserve assigns to a particular translator.
- Its rev org: company that has carried out the revision (Linguaserve).
- Its rev person: 20697, being this number the code Linguaserve assigns to a particular proofreader.

If the original file in German is modified and resent for translation, Linguaserve's localization and internationalization platform reads this information and preselects the same translator and proofreader when the project manager creates the batches for translation or proofreading, as it is show in the following screenshots:

Creation of translation batch:



Creation of revision batch:



4.3.5.1. ADVANTAGES

1) It provides useful information for project managers.

2) It promotes translation consistency.

4.3.5.2. DISADVANTAGES ENCOUNTERED

None.

4.3.5.3. CONCLUSIONS

As in the previous case, no disadvantages have been founded in the application of this metadata which reveals really useful for both management and consistency of translation and localization projects.

5. BUSINESS REPORT AND EXPLOITATION

After careful consideration VistaTEC decided to contribute Ocelot to Open Source. Since announcement Ocelot has been downloaded 54 times and has received much positive feedback on the Okapi Groups and Twitter. Phil Ritchie has spent much time during September and October doing outreach about Ocelot. Phil has given presentations at Machine Translation Summit in Nice, European Language Industry Association in Malta, Localization World in Santa Clara and TAUS DQF Conference in Santa Clara.

Many Language Service Providers has expressed a keen interest in Ocelot as the use case has clear benefits.

5.1. CMS-TMS use case

ITS 2.0 increases both user's control and automatic decision processes, allowing more "intelligence" in the systems involved:

- Translate: can be automatically or manually annotated. E.g. it allows not to add "non-translatable" terms in several specific glossaries or MT systems.
- Localization Note: direct communication between webmasters, PMs and translators. When alert type, it can be used for triggering certain processes in the Translation Workflow.
- Domain: automatic selection of CAT/MT terminology and dictionaries. Selection of Translation Memories by domains.
- Language Information: quality check to ensure the source language content is according to the Webservice parameter.
- Allowed Characters: quality check for the target content.
- Storage Size: quality check for both original content and the target content to be uploaded into the CMS. It can be used also for translators' visual control.
- Provenance: possibility to reassign the same translator/reviewer in new versions of the same content and inform the PM. Tracking control in the CMS.

- Readiness (ITS 2.0 extension): control of processes to be done, date control for availability, delivery and priority.

These and many other usage examples improve CMT-TMS systems in different aspects:

a) Management and process: the several mechanisms that ITS 2.o provides for a fine-grain communication between different actors, like webmaster and project manager, and this with translators and reviewers, as well as the use of "Readiness" makes more efficient the process and decrease s significantly the management costs.
b) On the other hand, while there is not a direct impact in translation cost, increases enormously the capacities to influence and improve quality and allows to externalize certain task s that has an final impact in the translation costs as well.
c) Finally, saving and reducing *a posteri* corrections and interventions, highly reduces the no-quality costs.

ITS 2.0 impact

5.1.1. SWOT ANALYSIS

In this section we will analyse the strength, weaknesses, opportunities and threats regarding different business perspectives.

Strengths
Management cost reduction: Faster and more efficient communication between localization chain actors. Specific information (notes, issues, etc.) can be coded in the content, allowing its automatic processing and fine grain treatment by project managers, translators, localization engineers, etc.
Increase of Translation Control: More efficient control over the content, ITS 2.0 allows a more dynamic and precise control internally to the content.
Independency of metadata from CMS technologies, localization platforms and formats
Delivery time reduction, reduction of communication and coordination between project management and translators and the clients, might have an effect of reducing delivery time.
Delivery time and process control, by using ITS 2.0 extensions (Readiness)
Quality Assurance: it allows new and better mechanisms for quality assurance

Weaknesses
Interoperability standardization: There is not a CMS-TMS interoperability standard. Readiness can simplify the web service standard, since part of the key information can be included in the content, and not in the Webservice or XML“wrap”. In case of redundancy, this could be used for quality checks.

Opportunities
Accelerator for adoption: of interoperability localization chains.
Improvement actual clients: for those clients already using it. For example, by measuring the non-quality costs and client’s project management reduction.
Increasing fully automatic processes and localization: expert systems.
SMEs: cannot afford proprietary complex developments and analysis and format specifications. Normalization also means availability.
Technological simplification: It simplifies interoperability implementations.
New methodologies, methods, and Language Technologies: Applying ITS 2.0 in CAT tools, post editing tools and Machine Translation, Translation Management Systems.
New standards: Interoperability and readiness information, and Quality Assurance.

Threats
Content creators adoption: Annotation of source content. This is probably the greatest threat, convincing content creators about the benefits of annotating the source content in the CMS.
CMS developers adoption: providing the content creators with suitable tools for automatic, semiautomatic, and manual annotation.
SEO applications: developing strategies to use ITS 2.0 metadata to improve content coherence for SEO and SEM

We can conclude that ITS 2.0 opens up ways of win-win business. It can make accelerating adoption of interoperability localization chains for new users, and resolving certain problem and generating quality and control improvements to existing users:

- More efficient control over the content and faster fine-grain communication between localization chain actors (e.g. webmaster/project manager)
- Localization platforms and format Independent:
- Better web and linguistic technology machine/machine interaction
- Better web and localization human/machine interaction
- Increasing fully automatic processes and localization expert systems in CMS and TMS.
- Opens up ways for connectors, pre- and post-editing, and CAT tools

The business opportunities are in very frequently updated web sites that need efficient multilingual updates and maximum control, such as corporate and industry information, e-Government, e-Commerce or Educational web sites. Also, ITS 2.0 can benefit those environments with highly distributed content creation through the CMS, the Web 2.0 and user content created(applying MT systems for immediacy) and using ITS 2.0 to contribute for multilingual SEO.

5.2. Beyond 2013:

The SWOT analysis and the needs of permeation of ITS 2.0 in the society suggest at least the following future actions.

5.2.1.1. EXTENSION READINESS

"Readiness" is not part of ITS 2.0 since, in the given time frame, the W3C MultilingualWeb-LT Working Group could not find consensus on all aspects of Readiness. The implementation of Readiness as an extension to ITS 2.0 allows to gather experience and to consider this data category for a future version of ITS. This approach already helped after the creation of ITS 1.0 to develop other data categories that are now a stable part of ITS 2.0. Linguaserve is applying Readiness in both use cases involved:

- Applied in CMS-TMS showcase (WP3)
- Applicability in Online Translation system (see D4.2.1 and D4.2.2)

It indicates the readiness of a document for submission to L10n processes or provides an estimate of when it will be ready for a particular process. It can be used in expert systems for automatic processing.

The current status of Readiness data model:

- ready-to-process – type of process to be performed next
- process-ref – a pointer to an external set of process type definitions used for ready-to-process
- ready-at – defines the time the content is ready for the process, it could be some time in the past, or some time in the future
- revised – indicates is this is a different version of the content that was previously marked as ready for the declared process
- priority – the priority of the content for the process
- complete-by –target date-time attribute for completing the process

In the TMS-CMS use case, Readiness is implemented as follows:

- TMS Engine: Update the data category information with the availability dates and the following tasks in the localization chain
- Localization workflow: Delivery date control and priority control

5.2.1.2. DISSEMINATION AND TRAINING

It is needed a strategy of awareness and evangelization of ITS 2.0 of key actors for widely adoption:

- Content creators and specially webmasters.
- CMS developers

- Final clients, specially SMEs since it is usually harder to reach them.

Some mechanism can be part of this strategy:

- Maintaining an active ITS 2.0 group as reference and expert group.
- Selection and continuous publishing in specialized magazines (best known content creation magazines, CMS magazines, etc.)
- Participation in events where these actors are.
- Creating a training program
- Exploitation and awareness of existing business cases (actual showcases) and new ones.

5.2.1.3. METHODOLOGIES AND TOOLS

In order to adequately using ITS 2.0, CAT tools, Translation Management Systems, post-edinting tools and Machine Translation engines need to improve their products.

Also, new methodologies and methods have to take place in the Industry and in the translation studies, as well as best practices for developers.

Other side-effect benefits need to be explored, such as possible applications of ITS 2.0 to improve content coherence for SEO and SEM.

5.2.1.4. STANDARDIZATION

New standards of interoperability can be empowered, using also ITS 2.0 extensions like Readiness. Quality Assurance standardization initiatives can be benefits from ITS 2.0 meta data (e.g. Localization Quality Issue, Localization Quality Rating or MT Confidence data categories).

6. GLOSSARY OF TERMS AND ACRONYMS

Term/Acronym	Definition
B2B	Business-to-business: electronic communications between businesses or enterprises.
CAT Tool	Computer Aided Translation tool.
CMS	Content Management System.
Drupal	An open-source Content Management System (CMS).
GBCC	Global Business Connector Contents (Linguaserve's software)
ITS	The Internationalization Tag Set (ITS) is a set of attributes and elements designed to provide internationalization and localization support in HTML5 and XML documents.
JQuery	A multi-browser JavaScript library designed to simplify the client-side scripting of HTML.
HTML	HyperText Markup Language.
L10N	Localization.
LSP	Language Service Provider
PLINT	Platform for Localization, Interoperability and Normalization of Translation (Linguaserve's software)
TMS	Translation Management System.
VDMA	Verband Deutscher Maschinen- und Anlagenbau (German Engineering Federation)
WYSIWYG	What You See Is What You Get
XHTML	eXtensible HyperText Markup Language.
XML	eXtensible Markup Language.
XPath	XPath is a language for addressing parts of an XML document, designed to be used by both XSLT and XPointer.

7. REFERENCES

Apache OFBiz:

http://ofbiz.apache.org/

Cocomore

www.cocomore.com

Drupal Official Web:

http://drupal.org/

ITS Interest Group:

http://www.w3.org/International/its/ig/

ITS 2.0 Requirements:

http://www.w3.org/TR/its2req/

ITS 2.0 Tag Set:

http://www.w3.org/TR/its20/

Linguaserve

www.linguaserve.com

MultilingualWeb official web:

http://www.multilingualweb.eu/

MultilingualWeb-LT EC Working Group:

http://www.w3.org/International/multilingualweb/lt/

SOAP specifications:

http://www.w3.org/TR/soap/

STAR Transit:

http://www.star-group.net/ENU/group-transit-nxt/transit.html

VDMA

www.vdma.org

VDMA Machines for Plastics:

http://www.machines-for-plastics.com/kug/

VistaTEC

www.vistatec.com

W3C XHTML 1.0 recommendation:

http://www.w3.org/TR/xhtml1/

W3C XML 1.0 recommendation

http://www.w3.org/TR/xml/

W3C XPath 1.0 recommendation

http://www.w3.org/TR/xpath/