Visualization of ITS 2.0 Metadata for Localization Process

Renat Bikmatov¹, Nathan Glenn², Serge Gladkoff¹, Alan Melby²

¹ Logrus International
² LinguaTech International

renatb@logrus.ru, nathan.g.glenn@gmail.com, sgladkoff@logrus.ru, akmtrg@byu.edu

Abstract

The Internationalization Tag Set (ITS) 2.0 specification was introduced by W3C in 2013 as a complement to the XML, HTML5 and XLIFF specifications. ITS 2.0 format can be used to exchange localization instructions and other context metadata between data processing tools, and also to deliver metadata to any person who is working on the content. Previewing ITS 2.0 metadata was the goal of this project. We present preview tools which enable translators and reviewers to refer to localization context information visually presented in a web browser window while working on the content in their content editor, CAT or other tool. The core ITS metadata categories, such as Localization Note, Terminology, and Translate, will help to bind the localization instructions to the content, and also to fill the gap between limited functionality of current CAT tools and required access to context information.

Keywords: ITS 2.0 metadata, metadata, context, preview, localization, localization instruction, translation, CAT, editing, machine translation, post-editing, XML, HTML, HTML5, XLIFF

1. Introduction

The postulates listed below in this section are based on the rich experience of Logrus as a localization company.

Localization in general is still suffering from the following unresolved problems: 1) previewing the source content in the final or publication format, and 2) supplying localization-related context information and instructions to translators and editors. With the increased use of CMS and a mass transition to asynchronous update and fragmented translation of content by bits and pieces, these problems are only becoming more severe. Linking glossaries, translation instructions and style guides to the source content presents another problem.

The source content is usually provided to translators as XML or XLIFF files. The source content provided for preview usually comes as raw XML without any support for its visualization. These formats are not easily readable; contextual information is often missing. Lack of mapping of terminology, trademarks, client instructions and other context information to the content to be localized has been identified as major cause of disruption of human work. The Internationalization Tag Set (ITS) was created to relieve these problems.

2. Existing Limitations on ITS Usage

In real world production environments, direct integration of ITS 2.0 or any other metadata into content to be localized is hampered by the following issues:

• Support of ITS 2.0 by available CAT and other authoring tools is either missing or limited.
• The separate implementations of ITS 2.0 by many CAT tools are sure to contain many discrepancies.
• Even if one day all authoring tools could fully support ITS 2.0 format, compatibility with legacy Translation Memories (TM) would still be an issue. New pieces of content enriched with ITS metadata markup would not fully match the same pieces without the metadata. As a result, you would not be able to reuse 100% matches, for example.
• Full support of ITS 2.0 metadata, local markup, global rules and external data by existing translation memory formats and engines still remains an open issue.
• ITS 2.0 representation in XLIFF format is still a work in progress. For more details, see:
  http://www.w3.org/International/its/wiki/XLIFF_Mapping/,

- At the moment, the content stored in CMS databases as XML is often converted to XLIFF for localization purposes using proprietary tools which control the segmentation of the content into translatable items (translation units). These tools do not use ITS metadata for the purpose of fragmentation. As a result, the ITS metadata embedded in the source XML files do not control the distinction between translatable and untranslatable content, for example. Moreover, many ITS metadata may be lost during conversion to XLIFF.

3. The Solution

The solution we propose is to separate data processing (localization) from previewing context information (visualization of metadata), while still providing some synchronization between the production and preview environments.

The solution developed by Logrus and LinguaTech, known as Work In Context System or simply WICS, implies generating a reference file (viewable source) that is provided to the translator/editor in addition to the pieces of source text to be translated (translatable source) using a CAT tool. The reference file is standard HTML5 containing the same ITS 2.0 metadata as the source file, and specialized CSS/JavaScript code is used to display the ITS metadata. This preview does not require additional proprietary or specialized software – any supported web browser is sufficient, making the solution very portable (see Fig. 1):

To support the previewing of XML and XLIFF files, text conversion utilities were developed to transform these formats into an HTML5 preview format with equivalent ITS 2.0 metadata. Mapping ITS metadata from XML to HTML format turned out to be non-trivial task. (See the documentation published at https://github.com/renatb/ITS2.0-WICS-converter/ for more details on the format conversion related issues.)

The conversion of source HTML files enriched with ITS 2.0 markup to preview-ready HTML files has been introduced for technical reasons to ensure that all ITS rules are gathered inside the document and no external rule files exist. It simplifies parsing rules and implementation of rule priority and inheritance logic.

When the viewable source is loaded in a browser, ITS 2.0 information is highlighted, color-coded and augmented with popups. Additional information contained in the ITS metadata is also shown, such as definitions, comments, instructions, parts of speech, semantic information, reference web sites (both extranet and intranet), reviewer’s comments, etc. Actual translation or editing might be carried out in another format in a CAT tool, but a parallel preview is certain to improve the view of the context for translators, editors, reviewers and other text workers in a wide variety of scenarios, including content authoring, translation, MT post-editing, knowledge transfer, etc.

See Figure 2 for an example of content enriched with ITS 2.0 localization metadata and rendered in a web browser. The content and localization metadata

---

**Figure 1. Content and metadata preparation for preview**
preview and navigation functionality is provided via JavaScript. There are two preview modes: brief and extended. In brief preview mode, the pieces of content linked to metadata are highlighted, but the metadata are not displayed. In extended preview mode, the metadata linked to any particular piece of content or several metadata items within the active fragment of content are displayed in a separate metadata preview panel in the browser window. This solution supports metadata nesting: you can assign some metadata to the inner part of a phrase even when there is already existing metadata assigned to that phrase.

- Ready to use JavaScript files and other auxiliary files automatically referenced by web browser when opening the preview-ready HTML files.
- The complete set of source code packages, auxiliary files and instructions on building all the project utilities.
- The extended project report for the end-users and solution developers.

The project deliverables have been published at GitHub as an Open Source project available under MIT license. See the references at the end of the paper.

4. Project Deliverables

The project deliverables include the following:

- Ready to use executables (CLI and GUI) of data converters used to transform XML, HTML or XLIFF files enriched with ITS 2.0 markup into preview-ready HTML files enriched with equivalent ITS 2.0 metadata.

5. Conclusions and Future Work

The main questions motivating this project were: how to preview ITS 2.0 metadata and how to use these metadata in real-life localization processes. This project sought to provide a portable ITS previewing solution, and to research the challenges associated with such previewing. With the tools we developed, localization instructions or other information can be shared effectively and
consistently with text workers via viewable versions of any files provided for reference including full versions of source files or any reference files enriched with ITS metadata, regardless of preferred platform or CAT tool.

With the development of the ITS 2.0 specification, the localization industry gained a carrier of localization metadata for major formats of content: XML, HTML, and XLIFF. One of the tasks for future work could be parsing and automatic or semi-automatic mapping of relevant rules from any stand-alone localization instructions to the relevant pieces of content via ITS markup. In an ideal case, such natural language processing solution should be able to apply ITS markup to the content according to any external localization instructions represented in some machine-readable format.

Acknowledgements

This work has been supported by Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) and carried out as a part of one of the tasks of the W3C LT-Web project (http://www.w3.org/International/multilingualweb/lt/), funded by the European Commission through the Seventh Framework Programme (FP7) under contract No. 287815.

References


