MLW-LT Working Group
Use Cases and Implementations

Version: October 2012

Simple Machine Translation

Description (brief description of use case. If more text is needed, add a second slide)

- XML and HTML5 documents are translated using a machine translation system, such as Microsoft Translator.
- The documents are extracted based on their ITS properties and the extracted content is send to the translation server. The translated content is then merged back into its original XML or HTML5 format.

Data Categories (list of categories used)

- Translate
- Locale Filter
- Element Within Text
- Preserve Space
- (Domain)

Benefits (list of anticipated benefits)

- The ITS markup provides the key information that drives the extraction in both XML and HTML5.
- Information such as preserving white space can also be passed on to the extracted content and insure a better output.

Simple Machine Translation

Detailed description of data category usage

- Translate - The non-translatable content is protected.
- Locale Filter - Only the parts in the scope of the locale filter are extracted, the others are treated as 'do not translate' content.
- Element Within Text - The information is used to decide what elements are extracted as in-line codes and sub-flows.
- Preserve Space - The information is passed on to the extracted text unit.
- (Domain) - The domain values are placed into a property that can be used to select an MT engine.

Translation Package Creation

Description (brief description of use case. If more text is needed, add a second slide)

- XML and HTML5 documents are extracted into a translation package based on XLIFF.
- The documents are extracted based on their ITS properties. The extracted content goes through various preparation steps and save into an XLIFF package. The ITS metadata passed on and carried by the extracted content are used by some steps.

Data Categories (list of categories used)

- Translate
- Locale Filter
- Element Within Text
- Preserve Space
- Id Value
- Domain
- Storage Size
- External Resource
- Terminology
- Localization Note
- Allowed Characters

Benefits (list of anticipated benefits)

- The ITS markup provide the key information that drives he extraction in both XML and HTML5.
- The documents to localize can be compared against older version of the same documents using ID to retrieve match the entries, and existing translations can be retrieved automatically.
- Information such as the domain of the content, external references, localization notes are available in the XLIFF document so any tool can make use of them to provide various translation assistance.
- Terms in the source content are identified and can be matched against a terminology database.
- Constraints about storage size and allowed characters can be verified directly by the translators as they work.

Translation Package Creation

Detailed description of data category usage

- Translate - The non-translatable content is protected.
- Locale Filter - Only the parts in the scope of the locale filter are extracted, the others are treated as 'do not translate' content.
- Element Within Text - The information is used to decide what elements are extracted as in-line codes and sub-flows.
- Preserve Space - The information is mapped to xml:space.
- Id Value – The value is mapped to the name of the extracted text unit.
- Domain – The values are placed into an <okp:itsDomains> element.
- Storage Size – The size is placed in maxbytes, and the native ITS markup is used for the other properties.
- External Resource - The URI is placed in a okp:itsExternalResource attribute.
- Terminology - The terminology information is placed into a specialized XLIFF note element.
- Localization Note - The text is placed into an XLIFF note.
- Allowed Characters - The pattern is placed in its:allowedCharacters.

Quality Check

Description (brief description of use case. If more text is needed, add a second slide)

- XML, HTML5 and XLIFF documents are read with ITS and loaded into CheckMate, a tool that performs various quality verifications.
- The XML and HTML5 documents are extracted based on their ITS properties, and their ITS metadata are mapped into the extracted content. The XLIFF document is also extracted and its ITS-equivalent metadata also mapped.
- The constraints defined with ITS are verified using checkMate.

Data Categories (list of categories used)

- Translate
- Locale Filter
- Element Within Text
- Preserve Space
- Id Value
- Storage Size
- Allowed Characters

Benefits (list of anticipated benefits)

- The ITS markup provides the key information that drives the extraction in both XML and HTML5.
- The set of ITS metadata carried in the files allows the three file formats to be handled the same way by the verification tool.

Quality Check

Detailed description of data category usage

- Translate - The non-translatable content is protected.
- Locale Filter - Only the parts in the scope of the locale filter are extracted, the others are treated as 'do not translate' content.
- Element Within Text - The information is used to decide what elements are extracted as in-line codes and sub-flows.
- Preserve Space - The information is mapped to the preserveSpace field in the extracted text unit.
- Id Value - The ids are used to identify the entries with an issue.
- Storage Size - The content is verified against the storage size constraints.
- Allowed Characters - The content is verified against the pattern matching allowed characters.

Leverage of Validator.nu Library

Description (brief description of use case. If more text is needed, add a second slide)

- Takes HTML5 with its- that converts it to XHTML with its: prefixes.
- Command-line tool uses a general HTML5 library to create the XML output
- More info at https://github.com/kosek/html5-its-tools

Data Categories (list of categories used)

- All data categories are converted

Benefits (list of anticipated benefits)

- Allows processing of HTML5 documents with XML tools.

HTML5 + ITS Markup Validator

Description (brief description of use case. If more text is needed, add a second slide)

- W3C uses validator.nu for experimental HTML5, but "its-" attributes are not valid HTML5 at present, generates errors.
- This version is updated by Jirka to allow use of new ITS attributes.
- More info at https://github.com/kosek/html5-its-tools

Data Categories (list of categories used)

- All data categories are validated

Benefits (list of anticipated benefits)

- Allows validation of HTML5 documents that include ITS markup.
- Catches errors in ITS markup for HTML5.
- Sets stage for HTML5+ITS validator at W3C.

CMS to TMS and Online MT System

Description (brief description of use case. If more text is needed, add a second slide)

- The contents are generated in a language service client side CMS, sent to the LSP translation server, processed in the LSP internal localization workflow, downloaded from the client side and imported into the CMS. Will use XML+ITS 2.0 as interchange format.
- More details at http://tinyurl.com/8woablr (still under review)

Data Categories (list of categories used)

1. Translate
2. Localization note
3. Domain
4. Language information
5. *Allowed Characters**
6. *Storage Size**
7. *Provenance Translation Agent**
8. *Provenance Revision Agent**
9. *Readiness***

* Pending Final Definition // ** Extension for CMS (out of ITS 2.0)

Benefits (list of anticipated benefits)

- Tighter workflow interoperability LSP-CMS-CLIENT
- Higher control of the content by the client, the localization chain and team:
  - Automatic (e.g. Translate)
  - Semiautomatic (e.g. Domain)
  - Manual (e.g. Localization

Online MT System Internationalization

Description (brief description of use case. If more text is needed, add a second slide)

- Illustrates how ITS allows a HTML5 Content Author to communicate instructions about the translation to MT Systems and a content editor via a Real Time Translation System (RTTS) connected to different MT Service Providers. Will use XHTML5 or HTML5 as format (dependency on the client).
- Detailed description http://tinyurl.com/92rtuqa (still under revision)

Data Categories (list of categories used)

1. Translate
2. Localization note
3. Language information
4. Domain
5. *Provenance Translation Agent**
6. *Provenance Revision Agent**
7. *Provenance Source Language**
8. *LocalizationQualityIssue**
9. *Readiness***

* Pending Final Definition // ** Extension for CMS (out of ITS 2.0)

Benefits (list of anticipated benefits)

- Improved control over translation actions via RTTS
- Improved control over what to translate and not to translate
- Improved domain-specific corpus selection and disambiguation
- Improved available information for post-editing

Using ITS for PO files

Description (brief description of use case. If more text is needed, add a second slide)

- Extends ITS benefits to PO files
- Implementation: ITS Tool

Data Categories (list of categories used)

- Preserve Space
- Locale Filter
- External Resource
- Translate
- Element within Text
- Localization Note
- Language

Benefits (list of anticipated benefits)

- ITS tool ships with a set of default rules for various formats and uses these for PO file generation

Browser-Based Review

Description (brief description of use case. If more text is needed, add a second slide)

- Unified browser-based review process that adds automation for translation review using text analytics
- Maps proprietary translation provenance from different CAT tools into a common format that can be interlinked
- Presentation http://tinyurl.com/8mafmqu

Data Categories (list of categories used)

- Standoff Provenance
- Loc Quality Issue

Benefits (list of anticipated benefits)

- Streamlines current idea of translation → review with duplication of work
- Simplifies data harvesting about review
- Improves audit and quality correction

Simple Segment Machine Translation

Description (brief description of use case. If more text is needed, add a second slide)

- Integration of SMT with CMS for simple segment translation
- More information http://tinyurl.com/8qt3uen

Data Categories (list of categories used)

- Domain
- Translate
- Language Information
- Translation Agent
- MT Confidence

Benefits (list of anticipated benefits)

- Reduces the need for human checking to ensure the correct content has been translated using the correct language pair.
- Improves the quality of machine translation by matching the training corpora of the SMT engine used as closely as possible to the type of text being translated.

HTML-to-TMS Roundtrip Using XLIFF

Description (brief description of use case. If more text is needed, add a second slide)

- Tools: CMS-LION and SOLAS
- Service-based architecture for routing localization workflow between XLIFF-aware components; automates workflow between ITS-aware components
- More information at http://tinyurl.com/8qsjs6z

Data Categories (list of categories used)

- Translate
- Domain
- Disambiguation
- Terminology
- Provenance

Benefits (list of anticipated benefits)

- Can modularize and connect any number of specialized (single-purpose) components.
- Reduces cost of configuring and monitoring performance of workflow spanning a variety of components

Using ITS in the CMS

Description (brief description of use case. If more text is needed, add a second slide)

- Make ITS 2.0 accessible in WCMS Drupal to end-users without localization experience
- Support localization workflow in the CMS

Data Categories (list of categories used)

- Disambiguation
- Domain
- Revision Agent
- Translation Agent
- Translate
- Localization Note
- (Readiness)

Benefits (list of anticipated benefits)

- Adds ability to apply ITS 2.0 local metadata through Drupal WYSIWYG editor
- Global ITS 2.0 metadata can also be set at content node level
- Content + ITS 2.0 metadata can be sent to and received from LSP (incl. automatic content re-integration)
- Storage of provenance metadata (e.g. revision and translation agents)

CMS-Level Interoperability Using CMIS

Description (brief description of use case. If more text is needed, add a second slide)

- Web-services system that supports improved localization for CMS using ITS rules via CMIS and open asynchronous change notification for CMIS. (Currently command line)
  Demo video: https://www.scss.tcd.ie/~lefinn/CMS-L10n-DemoVideo.mp4

Data Categories (list of categories used)

- Readiness
- Pass-through of others (document level)

Benefits (list of anticipated benefits)

- Provides referencing mechanisms (one rule to multiple docs and multiple rules to individual docs) and precedence for applying ITS in CMIS.
- Adds polling capability to CMIS

Enrycher - Jožef Stefan Institute
More detailed slides at http://tinyurl.com/8ej55vj and http://tinyurl.com/9lv637o

Disambiguation and text analyis annotation

Description (brief description of use case. If more text is needed, add a second slide)

- Tool: Enrycher
- Disambiguate fragments in the HTML input, marking them up with ITS2.0 **disambiguation** tags
- Mark the document that a certain **text analysis annotation** tool was used on the content
- Preserve HTML tree
- Could be used by CMS or MT preprocessing

Data Categories (list of categories used)

- Disambiguation
- Text analysis annotation

Benefits (list of anticipated benefits)

- The ITS markup provides the key information about what entities are mentioned
- Provides means for specific translation scenarios
- Provides means for for text-data integration scenarios

Disambiguation and text analyis annotation

Detailed description of data category usage

- **Disambiguation** – Marking up fragment of text that mention named entities with their identity references or class references
- **Text analysis annotation** – Mark up the fragment with the fact that it was processed by a particular tool