D3.1.2: XLIFF ROUNDTRIPPING PLUS XSLT FOR HIDDEN WEB FORMATS

David Filip, Milan Karásek, Jirka Kosek, Sean Mooney, Dave O'Carroll, et al.

Distribution: Public

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

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	3.1.2
Deliverable title:	XLIFF Roundtripping plus XSLT for Hidden Web Formats
Dissemination level:	PU
Contractual date of delivery:	30th September 2013
Actual date of delivery:	15th October 2013
Author(s):	David Filip, Milan Karásek, Jirka Kosek, Sean Mooney, Dave O'Carroll, et al.
Participants:	UL, Moravia, UEP
Internal Reviewer:	Cocomore
Workpackage:	WP3
Task Responsible:	UL
Workpackage Leader:	Cocomore

Revision History

Revision	Date	Author	Organization	Description
1	27/09/2013	David Filip, Milan Karásek	UL/Moravia	Compiling information and discussion
2	30/09/2013	David Filip	UL	Draft, to be worked on by Jirka Kosek
3	01/10/2013	Jirka Kosek	UEP	Penultimate Draft
4	24/10/2013	David Filip	UL	Revised Version

CONTENTS

1. Executive Summary

This deliverable is twofold. Moravia and UL have been working on XLIFF roundtripping of ITS 2.0 categories based on output of XLIFF generators such as OKAPI Framework Tikal and Trinity College Dublin CMS-L10n.

To ensure proper seeding and extraction of ITS 2.0 categories from deep web formats UEP has developed ITS aware stylesheets for Docbook, a widely used deep web format.

UL's roundtrip is based on the publicly accessible and to be open sourced platform SOLAS (Service Oriented Localisation Architecture Solution).

2. Introduction

This public deliverable describes the service oriented infrastructure of an ITS2.0 aware ecosystem of tools.

3. The Technology

3.1. SOLAS

SOLAS LOCCONNECT

This component is orchestrating the progression of the XLIFF roundtrip. In a sense it supports all ITS2.0 categories as far as they can be encoded in a valid XLIFF 1.2 or XLIFF 2.0 file.

It supports entry of the following categories via its PMUI:

- Domain

It displays the following ITS 2.0 categories in its Project Management Viewer:

- Translate
- Term
- Text Analytics

LocConnect was developed using the PHP programming language. It runs on an apache web server on Windows or Linux with a MySQL database for storing data. It has a RESTful API that can be used to access or modify data stored in the database.

LocConnect serves as the business orchestration unit in the overall SOLAS architecture. Files are uploaded though a PMUI site where the user provides some meta-data that can be used to produce a workflow for the file. LocConnect then makes the file available to the other tools that need to process it through its API.

LocConnect can be accessed at http://demo.solas.uni.me/locconnect/

SOLAS MT BROKER

The MT Broker is written in PHP and runs on an apache web server. Calls to the MT service providers are made using PEAR's HTTP Request2 library. It wraps calls to a number of language service providers including Bing and Moses which provide machine translations for the given file. The MT Broker then populates the XLIFF file with the alternative translations from the language provider. It supports the alt-trans element of XLIFF 1.2 and the matches module of XLIFF 2.0. It pulls its tasks from LocConnect using the LocConnect API.

The services are registered with metadata about their level of support for XLIFF and ITS, so that the MT Broker can decide in what way the MT services will be tasked and consumed. Based on the capabilities of the registered systems, the MT Broker or Mapper supports the ITS2.0 metadata categories in one of two ways:

USING SERVICES THAT SUPPORT THE ITS2.0 <-> XLIFF MAPPING

In that case the incoming XLIFF file is routed as is to the MT service and accepted back as delivered by the MT service. In this case the burden of supporting the ITS metadata categories is on the service provider rather than on the broker.

USING SERVICES THAT ARE NOT GENERALLY ITS AWARE

The XLIFF file is parsed for ITS 2.0 categories and these are interpreted and only suitable portions of the source content are sent to the MT service. The incoming XLIFF file is later populated with alternative translations based on the suggestions coming from the MT service, ITS metadata are added and modified as required.

In this scenario the broker interprets the

- Translate

metadata category

And is able to add

- Provenance
- MT confidence

metadata.

The MT Broker can be accessed at http://demo.solas.uni.me/mapper/

SOLAS LKR

The Localisation Knowledge Repository is written in PHP and runs on an apache web server. It uses a MySQL database to store data. It can be run as a standalone web service or as part of the SOLAS productivity suite. It can pull jobs from LocConnect using the LocConnect API.

LKR can read and modify the following ITS 2.0 metadata categories as mapped into XLIFF 1.2 and XLIFF 2.0 and back:

- Translate
- Term
- Text Analytics
- Domain

The LKR can be accessed at http://demo.solas.uni.me/lkr/author/

WORKFLOW RECOMMENDER

This component is basically a plugin providing the orchestration component LocConnect with the prescribed workflow information.

The Workflow Recommender is written in PHP running on an Apache web server. It uses PEAR's HTTP Request2 package to request resources from LocConnect. It generates a workflow based on the criteria provided by LocConnect and then sends an XLIFF file, enriched with the workflow information, back.

The workflow recommender can be accessed at http://demo.solas.uni.me/wfr/

TARGET POPULATOR

The target populator takes an XLIFF file with alternative translations and selects the most appropriate one to automatically populate the target. The selection is made using the match quality attributes and the provenance records. It is written in Dart compiled to Javascript and as such can run in any modern web browser (however, Google Chrome or Firefox are recommended).

Complex decision making plugins can be designed and connected.

The target populator can be accessed at http://demo.solas.uni.me/wfr/

TEXT ANALYTICS AND TERMINOLOGY BROKER

The text analytics and terminology broker is used to wrap text analytics web services. The only service currently invoked by this broker is Tilde's terminology enricher. The broker is written using Dart and compiled into Javascript so it can be run on any modern web browser with Javascript support (however, Google Chrome or Firefox are recommended).

This is analogical to the MT Broker, the TAT broker sends and receives XLIFF files to and from services that are aware of the ITS<->XLIFF mapping of the

Term

Text Analytics

ITS 2.0 categories.

Otherwise, the XLIFF encoding and injection of metadata can be done on behalf of the services that are not aware of the mapping.

The Text Analytics Broker can be accessed at http://demo.solas.uni.me/TA/web/

SOLAS EXTRACTOR AND MERGER

This component can be used for extraction of translatable content into valid XLIFF 1.2 files in cases the workflow initiator does not produce the XLIFF files on its own.

Based on the XLIFF version parameter, this tool can use either the Tikal extractor

XLIFF Extractor using http://www.opentag.com/okapi/wiki/index.php?title=Tikal

XLIFF Merger using http://www.opentag.com/okapi/wiki/index.php?title=Tikal

Or the XLIFF 2.0 Toolkit

https://code.google.com/p/okapi-xliff-toolkit/

However the development of the XLIFF 2.0 Toolkit has not yet been finalized, so this functionality is experimental and currently not capable of merging back.

3.2. Moravia MT Services

THE SERVICE

Because generally accessible MT services are not yet ITS2.0 aware, Moravia has provided an MT service component that processes valid XLIFF 1.2 and files with mapped ITS 2.0 information. Moravia also developed experimental support for XLIFF 2.0 files with mapped ITS 2.0 metadata.

Moravia's service is based on Moses, and M4Loc, which in turn uses the OKAPI Framework Tikal functionality.

Moravia exposed a RESTful service that receives valid XLIFF 1.2 files and returns them with translation candidates encoded as <alt-trans> elements, eventually valid XLIFF 2.0 files that are returned with translation candidates within the XLIFF 2.0 Translation Candidates module. This service is registered with SOLAS MT Broker but can be used standalone with any valid XLIFF 1.2 or XLIFF 2.0 file though direct upload and download. The RESTful interface allows for easy integration.

All changes that Moravia developed have been committed under LGPL to the M4Loc project.

For details of this solution see deliverable **D4.3: XLIFF Roundtripping Prototype based on M4Loc Work and Okapi Tools**

SUPPORTED ITS 2.0 METADATA

- Domain
- Term
- Text Analytics
- Translate
- MT Confidence

3.3. ITS 2.0 Roundtripping Support in DocBook XSL Stylesheets

DocBook is very popular format for producing structured documents that are later published on the Web. Support for all ITS 2.0 attributes and elements was added into DocBook schema by UEP and will be distributed as a part of a next official DocBook schema release.

Sources of DocBook+ITS 2.0 schema are available at https://github.com/docbook/docbook/tree/master/relaxng/schemas/dbits.

UEP also modified DocBook XSL stylesheets to pass-through ITS 2.0 metadata into resulting HTML code when HTML output is generated from the DocBook sources. This means that once ITS 2.0 metadata are entered into the content they are available also to tools operating directly on the web content.

Similar effort was also planned for DITA. However current version of DITA XML standard does not allow using namespaces for attribute extensibility. UEP is working closely with OASIS DITA TC to enable this functionality in the future versions of DITA XML standard. Meanwhile guidelines for using ITS markup inside current version of DITA XML are being developed by UEP.

4. WEBPAGES

4.1. Using ITS 2.0 in DocBook

http://xmlguru.cz/2013/05/docbook-and-its2

4.2. MLW-LT, XLIFF/MT Round-Tripping

http://mlwlt.moravia.com/mlwlt-web-test/Presentation.aspx

4.3. Other webpages

TCD CMS-L10N

CMS-L10N is not part of this deliverable, it is used as a complementary element triggering the roundtrips that are developed and delivered via this deliverable.

http://phaedrus.cs.tcd.ie/DocumentTracking/

TROMMONS

Currently only the SOLAS-Match component of the SOLAS platform is open sourced, the productivity components used as ITS2.0 and ITS2.0 <-> XLIFF Mapping reference implementations will be open sourced by mid November 2013. All above described SOLAS components will be distributed through TRF github repository under the LGPL.

http://www.therosettafoundation.org/trommons/technology-resources/

http://trommons.org/

https://github.com/TheRosettaFoundation/SOLAS-Match