



D3.2.1: OKAPI OCELOT

Phil Ritchie

Distribution: Consortium Internal Report

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

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	3.2.1
Deliverable title:	Report on VistaTEC deliverables
Dissemination level:	CO
Contractual date of delivery:	31 st March 2012
Actual date of delivery:	08/12/2013
Author(s):	Phil Ritchie, CTO, VistaTEC
Participants:	UL, DFKI
Internal Reviewer:	TCD
Workpackage:	WP1
Task Responsible:	@@@
Workpackage Leader:	Felix Sasaki

Revision History

Revision	Date	Author	Organization	Description
1	11/09/2013	Phil Ritchie	VistaTEC	Initial version.
2	08/12/2013	Phil Ritchie	VistaTEC	Final version.

CONTENTS

Document Information	2
Revision History	2
Contents	3
1. Executive Summary	4
2. Reviewer's Workbench Features.....	4
3. Reviewer's Workbench Benefits.....	4
4. Reviewer's Workbench User Interface	5

1. EXECUTIVE SUMMARY

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

Ocelot implements the Language Quality, MT Confidence 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. See "*Ocelot Open Source And Technical Details*".

2. OCELOT FEATURES

Ocelot has the following features:

- Reads/writes Language Quality Issue and Provenance ITS 2.0 data categories. Also reads MT Confidence data category.
- 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. 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.

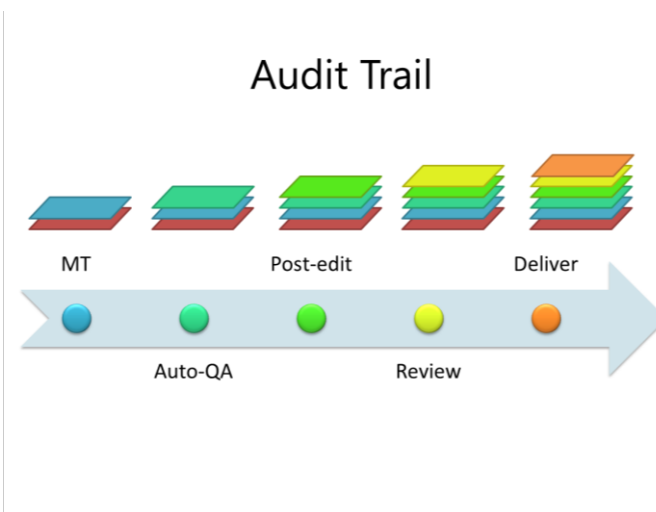
By way of a concrete example:

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.

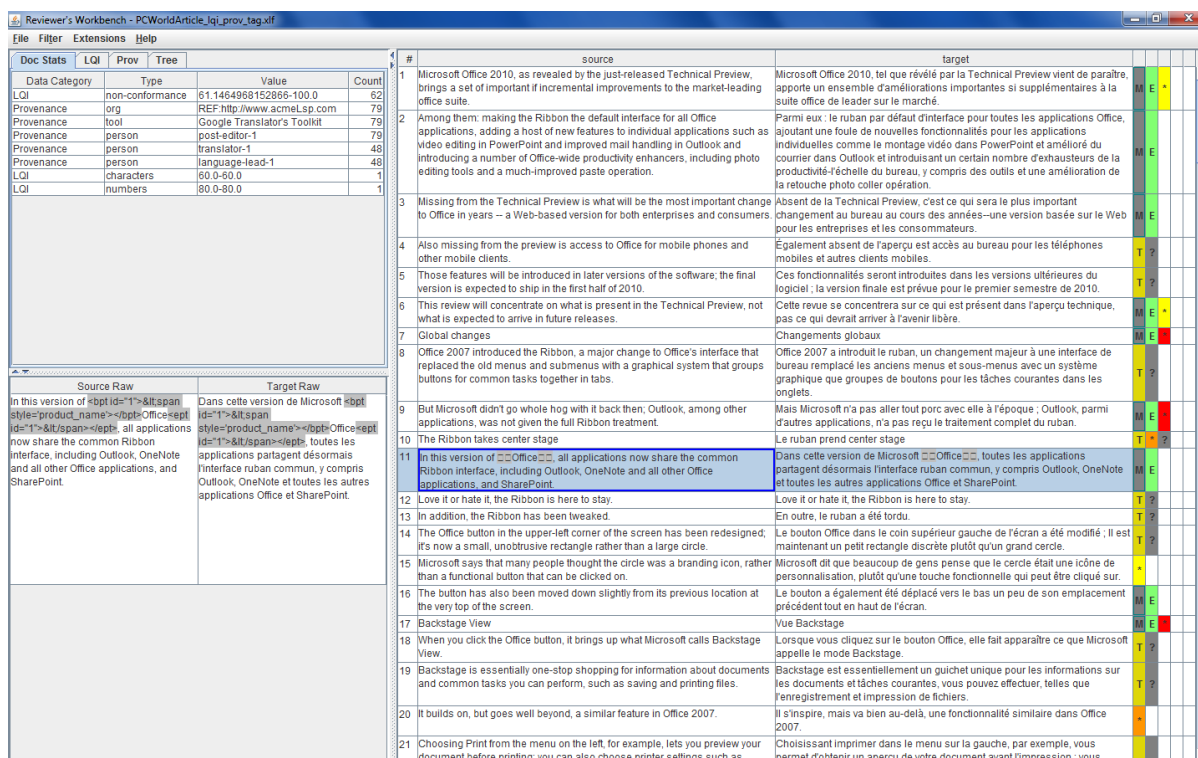
Following translation the document could be passed through an automated quality assurance pipeline. During this phase of the localization process the metadata listed below could be added to the segments:

- 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.

4. OCELOT USER INTERFACE



The screenshot shows the 'Reviewer's Workbench' window. The top-left panel displays a metadata summary table:

Doc Stats	LQI	Prov	Tree
Data Category	Type	Value	Count
LQI	non-conformance	61.1484968152866-100.0	62
Provenance	org	REF: http://www.acmel.ssp.com	79
Provenance	tool	Google Translator's Toolkit	79
Provenance	person	post-editor-1	79
Provenance	person	translator-1	48
Provenance	person	language-lead-1	48
LQI	characters	60.0-60.0	1
LQI	numbers	60.0-60.0	1

The main window displays a comparison table between source and target text:

#	source	target			
1	Microsoft Office 2010, as revealed by the just-released Technical Preview, brings a set of important incremental improvements to the market-leading office suite.	Microsoft Office 2010, tel que révélé par la Technical Preview vient de paraître, apporte un ensemble d'améliorations importantes et supplémentaires à la suite office de leader sur le marché.	M	E	-
2	Among them: making the Ribbon the default interface for all Office applications, adding a host of new features to individual applications such as video editing in PowerPoint and improved mail handling in Outlook and introducing a number of Office-wide productivity enhancers, including photo editing tools and a much-improved paste operation.	Parmi eux : le ruban par défaut d'interface pour toutes les applications Office, ajoutant une foule de nouvelles fonctionnalités pour les applications individuelles comme le montage vidéo dans PowerPoint et amélioré du courrier dans Outlook et introduisant un certain nombre d'exhausteurs de la productivité à l'échelle du bureau, y compris des outils et une amélioration de la retouche photo coller opération.	M	E	-
3	Missing from the Technical Preview is what will be the most important change to Office in years -- a Web-based version for both enterprises and consumers.	Absent de la Technical Preview, c'est ce qui sera le plus important changement au bureau au cours des années--une version basée sur le Web pour les entreprises et les consommateurs.	M	E	-
4	Also missing from the preview is access to Office for mobile phones and other mobile clients.	Egalement absent de l'aperçu est accès au bureau pour les téléphones mobiles et autres clients mobiles.	T	?	
5	Those features will be introduced in later versions of the software; the final version is expected to ship in the first half of 2010.	Ces fonctionnalités seront introduites dans les versions ultérieures du logiciel; la version finale est prévue pour le premier semestre de 2010.	T	?	
6	This review will concentrate on what is present in the Technical Preview, not what is expected to arrive in future releases.	Cette revue se concentrera sur ce qui est présent dans l'aperçu technique, pas ce qui devrait arriver à l'avenir libre.	M	E	-
7	Global changes	Changements globaux	M	E	-
8	Office 2007 introduced the Ribbon, a major change to Office's interface that replaced the old menus and submenus with a graphical system that groups buttons for common tasks together in tabs.	Office 2007 a introduit le ruban, un changement majeur à une interface de bureau remplacé les anciens menus et sous-menus avec un système graphique que groupes de boutons pour les tâches courantes dans les onglets.	T	?	
9	But Microsoft didn't go whole hog with it back then; Outlook, among other applications, was not given the full Ribbon treatment.	Mais Microsoft n'a pas aller tout porc avec elle à l'époque : Outlook, parmi d'autres applications, n'a pas reçu le traitement complet du ruban.	M	E	-
10	The Ribbon takes center stage	Le ruban prend center stage	T	?	
11	In this version of Office, all applications now share the common Ribbon interface, including Outlook, OneNote and all other Office applications, and SharePoint.	Dans cette version de Microsoft Office, toutes les applications partagent désormais l'interface ruban commun, y compris Outlook, OneNote et toutes les autres applications Office et SharePoint.	M	E	-
12	Love it or hate it, the Ribbon is here to stay.	Love it or hate it, the Ribbon is here to stay.	T	?	
13	In addition, the Ribbon has been tweaked.	En outre, le ruban a été tordu.	T	?	
14	The Office button in the upper-left corner of the screen has been redesigned; it's now a small, unobtrusive rectangle rather than a large circle.	Le bouton Office dans le coin supérieur gauche de l'écran a été modifié; il est maintenant un petit rectangle discrète plutôt qu'un grand cercle.	T	?	
15	Microsoft says that many people thought the circle was a branding icon, rather than a functional button that can be clicked on.	Microsoft dit que beaucoup de gens pense que le cercle était une icône de personnalisation, plutôt qu'une touche fonctionnelle qui peut être cliqué sur.	-		
16	The button has also been moved down slightly from its previous location at the very top of the screen.	Le bouton a également été déplacé vers le bas un peu de son emplacement précédent tout en haut de l'écran.	M	E	-
17	Backstage View	Vue Backstage	M	E	-
18	When you click the Office button, it brings up what Microsoft calls Backstage View.	Lorsque vous cliquez sur le bouton Office, elle fait apparaître ce que Microsoft appelle le mode Backstage.	T	?	
19	Backstage is essentially one-stop shopping for information about documents and common tasks you can perform, such as saving and printing files.	Backstage est essentiellement un guichet unique pour les informations sur les documents et tâches courantes, vous pouvez effectuer, telles que l'enregistrement et impression de fichiers.	T	?	
20	It builds on, but goes well beyond, a similar feature in Office 2007.	Il s'inspire, mais va bien au-delà, une fonctionnalité similaire dans Office 2007.	-		
21	Choosing Print from the menu on the left, for example, lets you preview your document before printing; you can also choose printer settings such as	Choisissant imprimer dans le menu sur la gauche, par exemple, vous permet d'obtenir un aperçu de votre document avant l'impression; vous	M	E	-

Figure 1: Main Workbench window. 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.

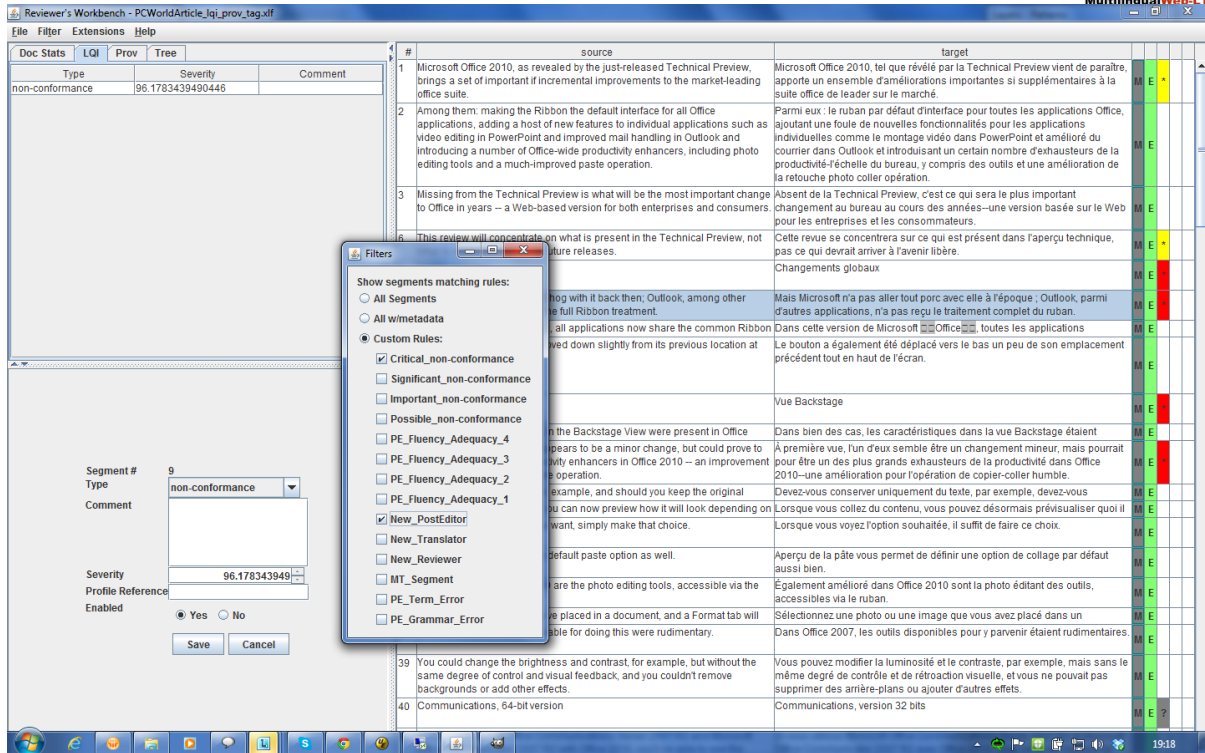


Figure 2: Popup window shows rules associated with particular metadata characteristics used as filtering rules.

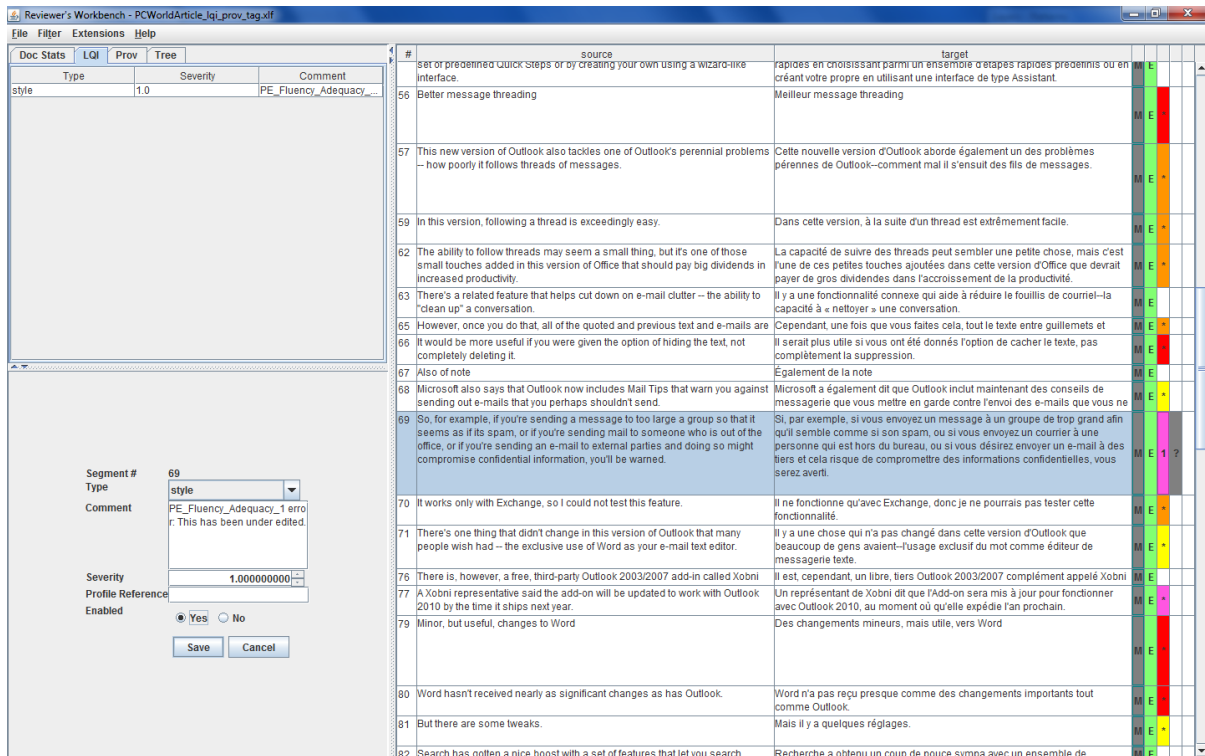


Figure 3: Bottom left-hand panel displays the Language Quality Issue data category values associated with the selected segment in the main editor window.

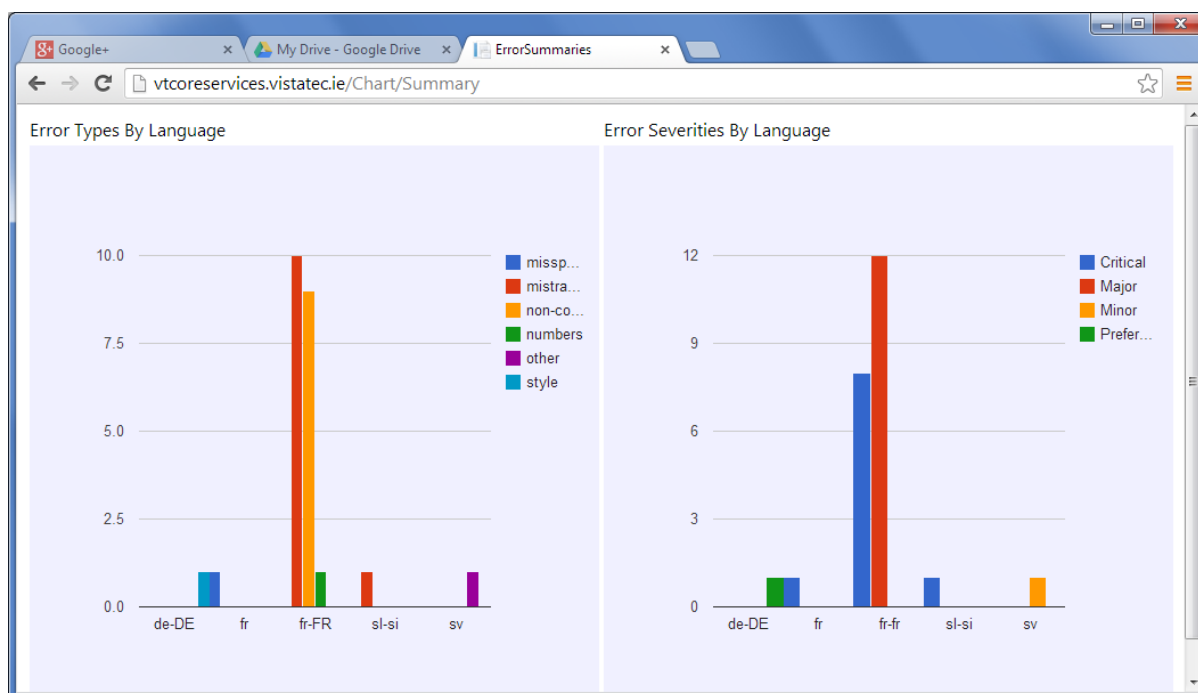


Figure 4: Quality Web Dashboard. LQI metadata can be sent via RESTful API to VistaTEC endpoint and quality metrics immediately updated in portal dashboard.

5. OCELOT OPEN SOURCE AND TECHNICAL DETAILS

VistaTEC's announced the availability of Ocelot as Open Source under the LGPL License at Localization World, Santa Clara in October 2013.

Source code is available at: <https://github.com/ocelot>

Documentation wiki is at: <http://open.vistatec.com/ocelot>

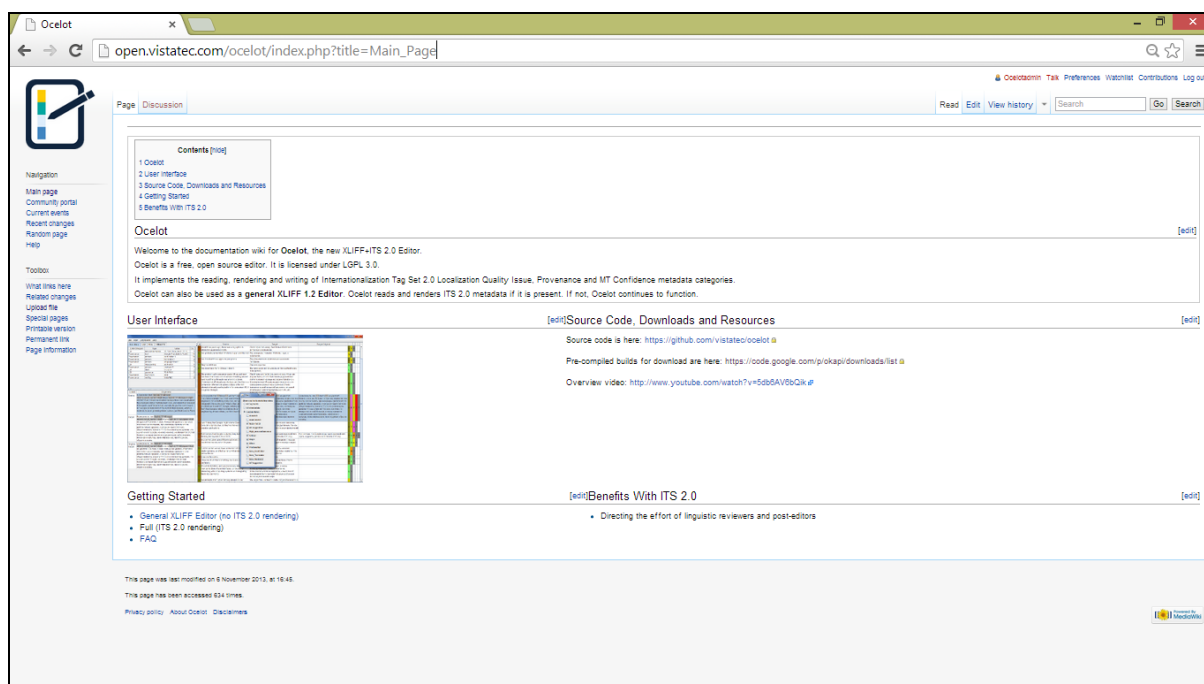
Binaries are available at:

Email list: okapitools@yahoogroups.com

Ocelot is written in Java.

It uses (and enhanced) several Okapi Framework classes, namely: `XLIFFFilter`, `XLIFFSkeletonWriter`, `ITSStandoffManager` and `ITSProvenance*`.

The easiest way of getting setup with the source code is to install the Oracle Netbeans IDE. Download the code from GitHub and open the Maven `pom.xml`. This will initiate Netbeans downloading all of the dependency libraries. Once this completes you can build and launch the application.



6. TANGIBLE OUTCOMES OF OCELOT

The technology startup Digital Linguistics (<http://www.digitallinguistics.com>) has adopted ITS 2.0 as its mechanism for writing its text analytics-generated Conformance Scores into XLIFF.

Ocelot and ITS 2.0 were presented at the TAUS Translation Quality Evaluation Summit held in Adobe on 9th October 2013. This event was attended by many of the world's top localizers.

Ocelot now forms the basis of a post-editing service within VistaTEC which is based fundamentally upon ITS 2.0 metadata.

VistaTEC's process now consists of the following steps:

1. Source files are converted to XLIFF 1.2. This can be done with Okapi Rainbow.
2. Files are sent to machine translation engine.
3. Files containing raw machine translation output are run through Okapi Checkmate. This adds ITS metadata for common errors.
4. Files are posted to Digital Linguistics' Review Sentinel service which adds "non-conformance" ITS markup.
5. Ocelot and a `rules.properties` file is distributed to post-editors. The `rules.properties` file contains rendering and segment filtering rules for 4x 10-point quality non-conformance bands. It also configures keyboard shortcut keys for adding Fluency/Adequacy scores.
6. Files are post-edited. Post-editors carry out edits using the 4 number 10-point non-conformance filters as their guide for segments to post-edit. Editors also add Fluency/Adequacy scores according to

proprietary guidelines.

7. Files are post-processed to harvest metadata added during post-editing.
8. Files are converted back to native format.

7. FUTURE RESEARCH AND COLLABORATIONS

Ocelot's currently unique position as the only Open Source editor to support XLIFF and ITS puts it in a good position to be the platform of choice for future research initiatives to implement their ideas and standards and avoid the programming overhead of writing the underlying XLIFF parsing, rendering and editing functions of an editor.

VistaTEC is an active member of communities who are progressing research in the area of Quality Estimation such as QTLaunchPad Multi-dimensional Quality Metrics and TAUS Dynamic Quality Framework.

Ocelot could easily be adapted to capture, use and generate semantic mark-up such as RDF, NIF, PROV and GLOBIC models.

VistaTEC is also a member of CNGL II and is supporting proposals for ADAPT (CNGL III). VistaTEC intends promote ITS within these forums.