



# Everything you wanted to know about the Internationalization Tag Set - ITS2.0

**Dave Lewis**

W3C MultilingualWeb - Language Technology  
Working Group

CNGL, Trinity College Dublin





## Internationalization (I18n) Activity

*Making the World Wide Web world wide!*

[Home](#) [About](#) [Groups](#) [Topics](#) [Techniques](#) [Resources](#) [News](#)

Internationalization site search:

RSS Feeds

## MultilingualWeb-LT Working Group Home Page

**Key WG information:** [Mission](#) - [Funding](#) - [News](#) - [Work in progress & Events](#) - [Charter](#) - [Archives](#) - [Working Group wiki](#) - [Participants](#) - [Contribute](#) - [Join](#)



Any links on this page to W3C member confidential pages are marked with <sup>MO</sup>.

### About the MultilingualWeb-LT Working Group

The MultilingualWeb-LT (Language Technologies) Working Group is part of the [W3C Internationalization Activity](#) and the [MultilingualWeb](#) community.

**Aims:** define the [Internationalization Tag Set \(ITS 2.0\)](#), that is: meta-data for web content (mainly HTML5) and "deep Web" content, for example a CMS or XML files from which HTML pages are generated, that facilitates its interaction with multilingual technologies and localization processes.

For more details, see the [Working Group Charter](#).

### Mailing Lists

The working group uses the [public-multilingualweb-lt](#) mailing list. [\[Archive\]](#) [\[Subscribe\]](#)

There is also a member-only list [\[Archive<sup>MO</sup>\]](#) which is used for group administravia and any

### Work in Progress & Events

A **second last call working draft of ITS 2.0** has been published 21 June 2013. The Working Group believes that this draft is feature complete - **provide feedback** by 11 June!

A **ITS 2.0 showcase** will demonstrate ITS 2.0 implementations in Dublin - [register now!](#)

The working group will produce **ITS 2.0**, the successor of [ITS 1.0](#). ITS 2.0 will be a W3C Recommendation about metadata for multilingual language technologies and localization processes.

A basis of the working draft for ITS 2.0 are the [requirements gathered for ITS 2.0](#). We [welcome feedback](#) on both the requirements document and the [ITS 2.0 draft](#).

Please see the [Working Group wiki](#) for further details on work items, face-to-face and virtual [meetings](#) etc.

### Press

### Contribute

We have gathered input at the [March 2012 MultilingualWeb workshop](#) in Luxembourg and at the [June 2012 MultilingualWeb workshop](#) in Dublin (see the [workshop report](#)).

Also, we conducted a dedicated questionnaire about international Web content requirements. We will use this questionnaire to gather initial input and expressions of interest in participation in our work.

We also welcome feedback on our [public mailing list](#) about

- the ITS 2.0 draft ([latest published version](#) and [editor's copy](#))
- our [Requirements and Use Cases](#) draft document
- [technical issues](#) discussed in the group, and
- the work plan as described in the [charter](#).

### Join

# What is ITS 2.0?



- Meta-data annotations addressing gaps in I18n and L10n interoperability
  - CMS, L10n Tools, Language Technology
- Defines 18 separate Data Categories
- Implementation-neutral specification of data categories
- Implementations in HTML5, XML and RDF

# ITS 2.0 Draft Data Categories

## ITS1.0

- Translate
- Localization Note
- Terminology
- Directionality
- Lang info
- Element within text

## I18n

- Locale Filter
- External Resource
- Preserve Space
- Allowed Characters
- Storage Size
- ID Value

## Language Technology

- Domain
- MT confidence
- Text Analysis

## Provenance & QA

- Quality Issue
- Quality Rating
- Provenance

# Who Benefits?

- LSPs, tools vendors, clients, CMS vendors, system integrators
- Industry-academia collaboration, EC support
- Implementing Partners:



# ITS2.0 by Numbers

41 WG  
Members

3500+ Posts  
to WG List

928  
successful  
Conformance  
Tests

9 Face-to-Face  
WG Meeting

57 online WG  
Meetings

15 Organisations  
Implementing

93 Examples  
in Spec

225 Test  
Suite Files

# Scope of ITS2.0 Use Cases









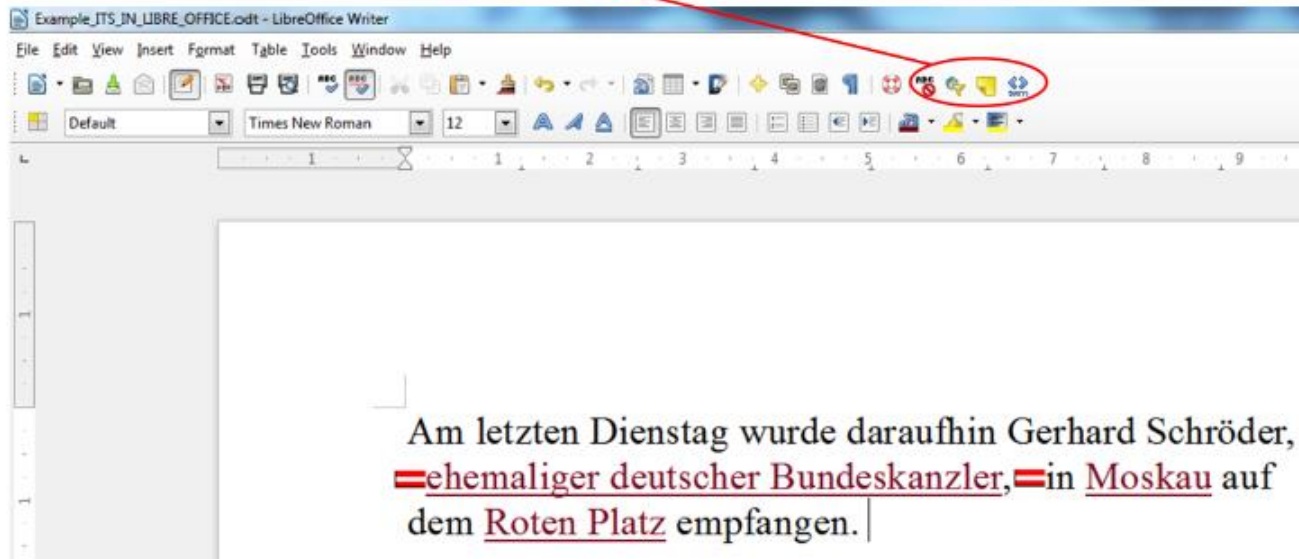


### ITS for text processing

The ]init[ ITS Libre Office Writer Extension "ITS for LO" showcases the use of a subset of the Internationalization Tag Set (ITS) in an open source word processing environment (Libre Office Writer)

### Supported ITS data categories

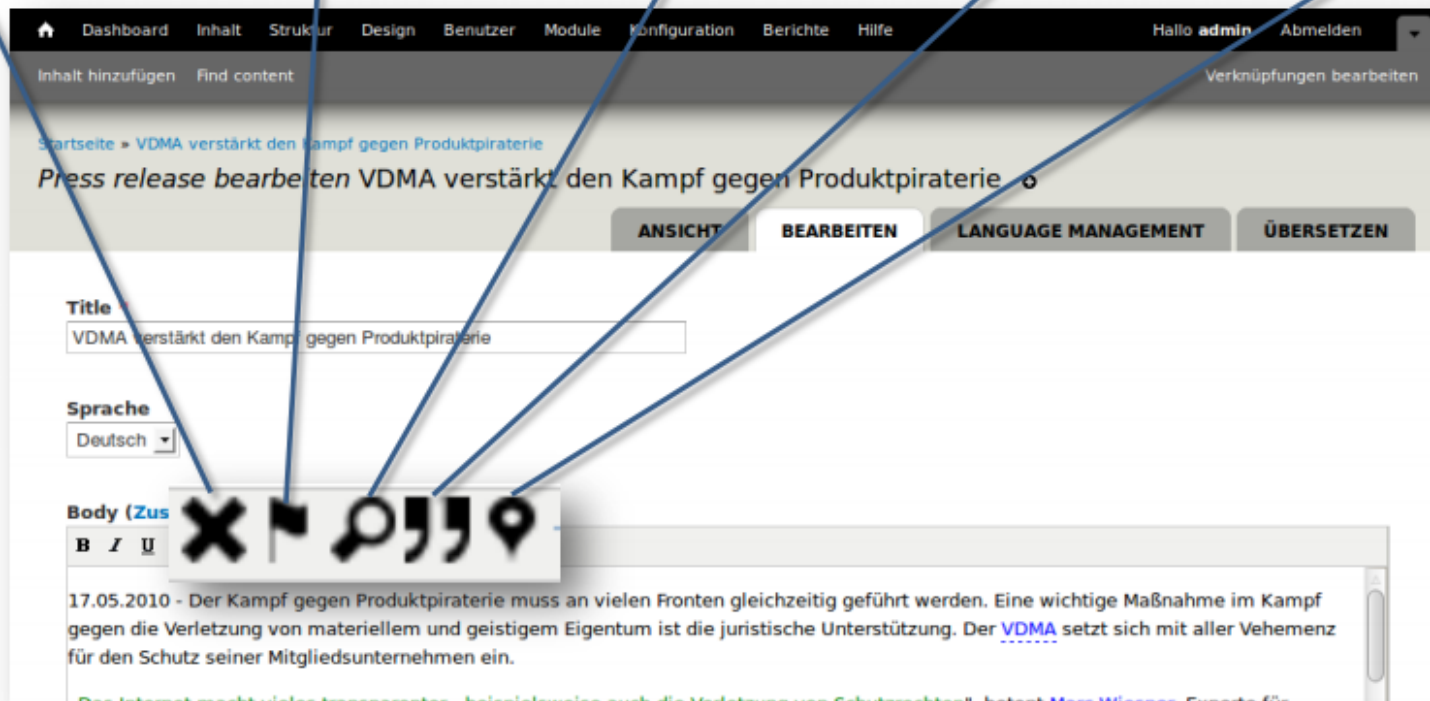
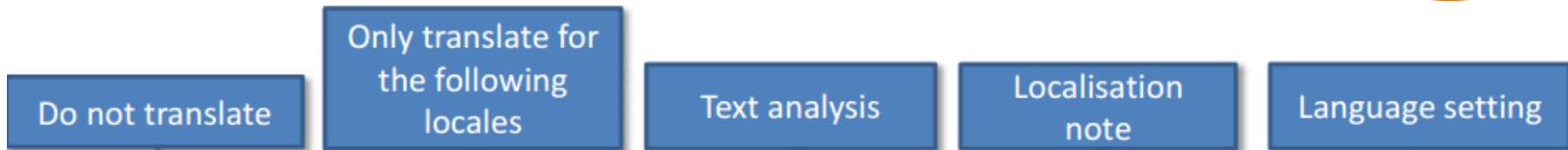
-  Terminology - One or several words can be marked up as "term"
-  Translate - will be used locally to set words as "not to translate"
-  Localization note - will be used to pass a note to the LSP
-  Locale Filter - will be used to limit phrases to specific locales



# CMS-based I18n



Content Management





Home

About

Contact



Home Events & Reports Blog About Partners & Contacts Useful links

# ITS 2.0 Enriched Terminology Annotation Showcase

## About

Given the fundamental nature of the World Wide Web for communication in all parts of the world, and as the share of English continues to decline and the number of languages spoken in the European Union continues to rise, the importance of ensuring the multilingual usability of the World Wide Web is paramount.

In order to build on current initiatives already in place and move it forward, it is necessary to take awareness of existing best practice and see what is still missing. It is important to be able to gather information from the various different countries that are involved in existing initiatives, to avoid needless duplication of work, and to ensure that initiatives are harmonised and interoperable.

The work of the MultilingualWeb initiative is being driven by the World Wide Web Consortium (W3C), an organization of around 400 member organizations worldwide that research and develop the Web's invention, Sir Tim Berners-Lee. It is supported with funding from the European Commission.

Welcome to the Terminology Annotation Showcase. The showcase allows you to automatically annotate terminology in [ITS 2.0](#) enriched content in HTML5, XLIFF and plaintext formats.

[ITS 2.0](#) content analysis and terminology annotation are performed by a dedicated Terminology Annotation Web Service API – a Web-based interface for statistical term candidate annotation using state-of-the-art methods as well as term bank based term candidate annotation using terminology resources from [EuroTermBank](#).

The Terminology Annotation Web Service API can be integrated in various natural language processing workflows, for instance, machine translation, localization, terminology management and other tasks that may benefit from automatic terminology annotation.

Select one of the options below to annotate terminology in plaintext, HTML5 or XLIFF content.

 Text

 HTML5

 XLIFF



# Named Entity Recognition



Terminology

- <http://enrycher.ijs.si/mlw/>

## enrycher

THIS IS WHAT YOU'RE WRITING ABOUT:

```
<!DOCTYPE html>
<div
  its-annotators-ref="text-analysis|http://enrycher.ijs.si/mlw/toolinfo.xml#enrycher">
  Hello, <span
    its-ta-ident-ref="http://dbpedia.org/resource/Rome"
    its-ta-class-ref="http://schema.org/Place">Rome</span>!
</div>
```

- Source code: <https://github.com/tadejs/enrycher-its20>

# Multilingual Publishing: Spanish Tax Agency

Translation

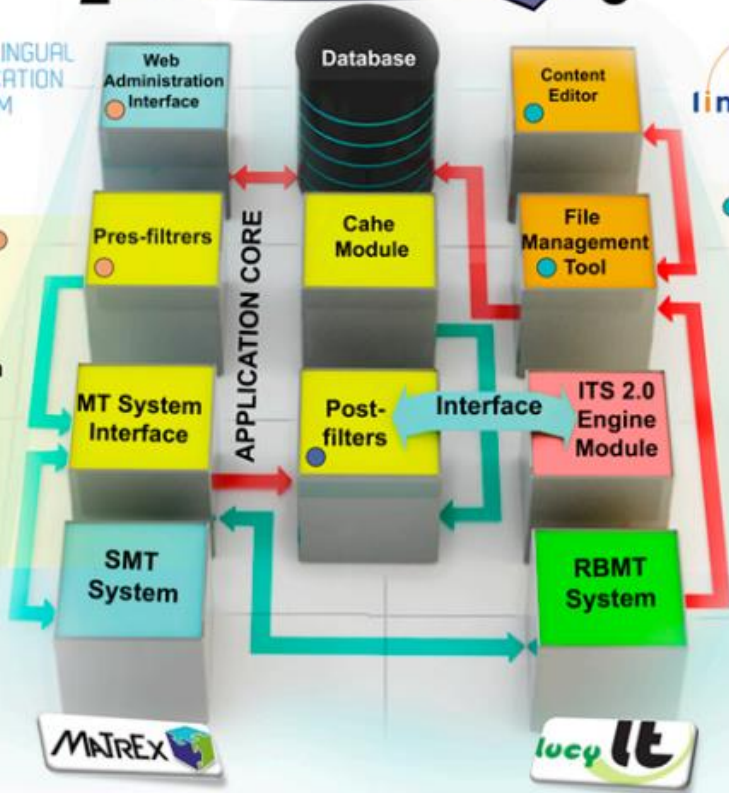


MULTILINGUAL PUBLICATION SYSTEM



**Pre-Production/Pre-filters**

- Translate** ► Adds constant marks to block non-translatable text nodes and attributes.
- Domain** ► Generates a HTML section per domain, and adds such information to the section sent to the MT System to select the appropriate TM and vocabulary (glossary).
- Localization Note** ► Encoding of information to convey it to the post-editor.
- Localization quality issue** ► Encoding of information to convey it to the post-editor.



**Content Editor + FMT:**

- Localization Note** ► Blocked by the Content Editor but visible for the post-editors.
- Domain** ► Stores the revised texts in the domain specific TM.
- Provenance** ► Adds and blocks the information about the reviser
- Localization Quality Issue** ► Adds and blocks the quality issue inserted by the reviser.

**Post-production/Post-filters:**

- Language Information** ► Updates the language attributes in the translated content.
- Provenance** ► Adds and blocks the information about the MT System



Transit XV [SP 26] - MLW-LT tests (User)

File Edit Cursor View Project Options Window Help

1 2 3 4 5

German: 1214\_193\_node\_norm

.....<xlasRefTrad url=""><![CDATA[http://staging-mlwlt.cocomore.com:8080/de/node/193]]></xlasRefTrad>  
 .....<xlasRefTrad it2dataCategory="Domain"><![CDATA[Allgemein, Angewandte Wissenschaft, Anlagenbau, Kunststoff- und Gummimaschinen, Maschinenbau, Technologien, Unternehmen]]></xlasRefTrad>  
 .....<xlasRefTrad it2dataCategory="LocalizationNote"><![CDATA[Pressemitteilung]]></xlasRefTrad>  
 .....<xlasTrad fieldName="36672-node\_title" xMax="255" xMaxOrig="48" xlasId="1"><![CDATA[«1+»  
 VDMA•verstärkt•den•Kampf•gegen•Produktpiraterie«2»  
 ]]></xlasTrad>  
 .....<xlasTrad fieldName="36672-body-0-value" xlasId="2"><![CDATA[«3+»  
 17.05.2010--Der•Kampf•gegen•Produktpiraterie•muss•an•vielen  
 Fronten•gleichzeitig•geführt•werden. •Eine•wichtige•Maßnahme•im•Kampf•gegen•die•Verletzung•von•materiellem•und•geistigem•Eigentum•ist•die•juristische•Unterstützung.  
 Der <xlasBlok xlasBloqId="1">&lt;span translate="no" &gt;VDMA&lt;/span&gt;</xlasBlok>•setzt•sich•mit•aller•Vehemenz•für•den•Schutz•seiner•Mitgliedsunternehmen•ein.«4»  
 </p>  
 <p>«5+»  
 „<span its-loc-note="Bitte korrekte sinngemäße Übersetzung mit Marc Wiesner absprechen." its-loc-note-type="description">Das•Internet•macht•vieles  
 transparenter••beispielsweise•auch•die•Verletzung•von•Schutzrechten</span>“,•betont•<xlasBlok xlasBloqId="2">&lt;span translate="no" &gt;Marc Wiesner&lt;/span&gt;</xlasBlok>•  
 Experte•für•Produktpiraterie•der•Abteilung•Recht•im•<xlasBlok xlasBloqId="3">&lt;span translate="no" &gt;VDMA&lt;/span&gt;</xlasBlok>•Viel•schneller•als•früher  
 bemerken•es•die•Unternehmensvertreter•heutzutage,•wenn•Produkte•angeboten•werden,•die•den•eigenen•täuschend•ähnlich•sind•oder  
 illegale•Nachahmungen•darstellen. •Das•weltweite•Datennetz•hilft•nicht•nur•beim•Verkauf•illegaler•Waren,•es•bringt•ebenso

French: 1214\_193\_node\_norm

.....<xlasRefTrad it2dataCategory="Domain"><![CDATA[Allgemein, Angewandte Wissenschaft, Anlagenbau, Kunststoff- und Gummimaschinen, Maschinenbau, Technologien, Unternehmen]]></xlasRefTrad>  
 .....<xlasRefTrad it2dataCategory="LocalizationNote"><![CDATA[Pressemitteilung]]></xlasRefTrad>  
 .....<xlasTrad fieldName="36672-node\_title" xMax="255" xMaxOrig="48" xlasId="1"><![CDATA[«1+»  
 VDMA•renforce•la•lutte•contre•le•piratage•des•produits«2»  
 ]]></xlasTrad>  
 .....<xlasTrad fieldName="36672-body-0-value" xlasId="2"><![CDATA[«3+»  
 17/05/2010--La•lutte•contre•le•piratage•des•produits•doit•être•menée•sur•de•nombreux•fronts•à•la•fois. •L'assistance•juridique•est•une•mesure•  
 essentielle•dans•la•lutte•contre•la•violation•de•la•propriété•matérielle•et•intellectuelle.  
 <xlasBlok xlasBloqId="1">&lt;span translate="no" &gt;VDMA&lt;/span&gt;</xlasBlok>•milite•avec•véhémence•pour•la•protection•de•ses•sociétés•membres.«4»  
 </p>  
 <p>«5+»  
 « <span its-loc-note="Bitte korrekte sinngemäße Übersetzung mit Marc Wiesner absprechen." its-loc-note-type="description">Internet•donne•de•la•transparence•à•beaucoup•de•choses, •  
 dont•notamment•à•la•violation•du•droit•d'auteur</span> »,•souligne•<xlasBlok xlasBloqId="2">&lt;span translate="no" &gt;Marc Wiesner&lt;/span&gt;</xlasBlok>•spécialiste•en•piratage•  
 des•produits•au•sein•du•département•juridique•de•<xlasBlok xlasBloqId="3">&lt;span translate="no" &gt;VDMA&lt;/span&gt;</xlasBlok>•De•nos•jours, •les•représentants•des•entreprises•  
 remarquent•beaucoup•plus•rapidement•qu'autrefois•la•ressemblance•frappante•entre•les•produits•proposés•et•les•leurs•ou•les•contrefaçons•illégalles. •Même•si•le•réseau•mondial•de•  
 données•aide•à•la•vente•de•marchandises•illégalles,•il•met•également•au•grand•jour,•rapidement•et•en•tout•lieu,•des•violations•du•droit.«6»  
 </p>

Line:16 Col:3 Seg:6 Pos:122 Ins Not translated D : Alt+068

ADOBE® CAPTIVATE™

Translation

# Spanish Tax Agency: QA Example

The quality Issue is fixed

Special Tax Return (STR)  
What is it used for and who has to declare the STR.

<p>What is used for and who has to <span its-loc-quality-issue-comment="Misspelling error, should be 'declarar'" its-loc-quality-issue-type="characters" its-loc-quality-issue-severity="90" its-loc-quality-issue-enabled="no">declare</span> the <span its-loc-note="Stands for 'Declaración Tributaria Especial', use acronym in target language" its-loc-note-type="description">STR</span></p>

# QA: Reviewers Workbench

VistaTEC

Quality Assurance

Reviewer's Workbench

File Filter Help

Doc Stats	LQI	Tree	+
Data Category	Type	Value	Count
LQI	other	25.0-25.0	1
LQI	style	10.0-25.0	14
LQI	mistranslation	25.0-75.0	3

#	source	target				
1	New firmware version (V4.00) that allows recording 720/50p in HQ mode (MENU-REC FORMAT)	Nuova versione di firmware (V 4.00) per registrazioni a 720/50p in modalità HQ (MENU-REC FORMAT)	\$	\$		
2	Ludwig Elektro- und Netzwerktechnik equips the newly opened EURONICS XXL specialist store in Hassfurt with a Sony video surveillance solution.	Ludwig Elektro- und Netzwerktechnik equipaggia EURONICS XXL, il negozio specializzato aperto di recente ad Hassfurt con una nuova soluzione di videosorveglianza Sony.	\$			
3	Scandinavian award for the VPL-VW1000ES Best Product of the Year	Premio scandinavo per il VPL-VW1000ES Miglior prodotto dell'anno	\$			
4	If all that puts you in the mood for ground-breaking	Se tutto ciò ti va venire voglia di saperne di più	m	m	\$	
5	We're the unrivalled leader in 4K digital cinema, raising the bar for unrivalled picture quality that has to be seen to be believed	Siamo leader incontrastati nel digital cinema 4K, elevando gli standard per creare una qualità delle immagini senza precedenti.	\$			
6	Great for budget-conscious communicators, they're ideal for medium-sized meeting areas, boardrooms and training suites.	Ideali per i comunicatori con budget limitati, sono ideali per le aree dedicate alle riunioni di medie dimensioni.	\$	m		
7	Sony 4K projection adds sparkle to Marrakech film festival	La proiezione 4K di Sony fa aumentare il prestigio del festival del cinema di Marrakech	\$	\$		
8	Sony builds on 'Year of Display Innovations' vision at ISE	Sony consolida ad ISE la sua visione "Year of Display Innovations"	o			
9	Super 35mm 4K CMOS sensor compact CineAlta camera records HD/2K/4K on SxS memory plus 16-bit RAW 2K/4K output	Telecamera CineAlta compatta con sensore CMOS 4K Super 35 mm, funzionalità di registrazione HD/2K/4K su memory card SxS e uscita RAW 2K/4K a 16 bit	\$			
10	said committee Chairman, Robert P. Seidel, Vice President of CBS Engineering and Advanced Technology.	ha dichiarato Robert P. Seidel, presidente della giuria e vicepresidente di CBS Engineering and Advanced Technology.	\$	\$		
11	So we've asked a few experts to tell us how these smaller investments can make a big difference to your footage	Ecco perché abbiamo chiesto ad alcuni esperti di offrirci il loro punto di vista su come questi piccoli investimenti possono fare una grande differenza per le riprese.	\$			
12	Camera accessories are as important as the camera itself when it comes to the smooth running of a shoot.	Gli accessori di una telecamera rivestono la stessa importanza della stessa telecamera nella riuscita delle riprese.	\$			

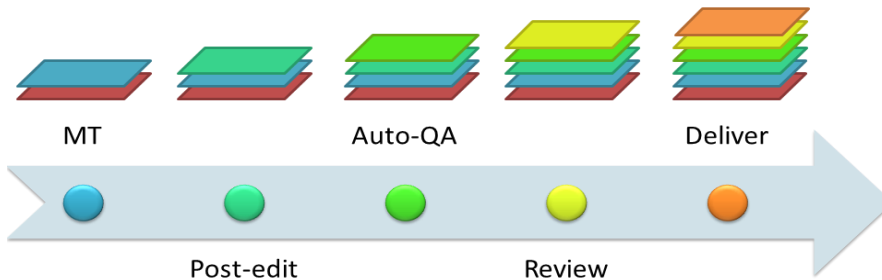
  

Segment #	6
Type	style
Comment	Repetition of "ideale" due to literal translation.
Severity	25.0
Profile Reference	
Enabled	yes



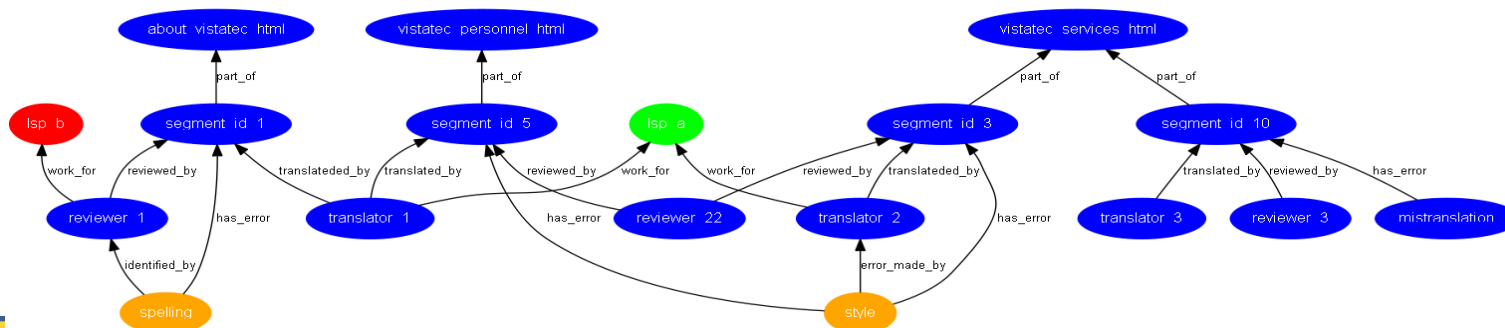
# Workflow/Resource Management

Workflow /  
Resource  
Management



VistaTEC

- Better up-front and real-time business decision support:
  - Use historic and aggregated metadata for inference
  - Content carries its own “audit trail”
- Metadata can be transformed to linked semantic data allowing:
  - Inferential queries across data siloes
  - New forms of data visualization
  - Extensible data relationships



# Summary

- Spec in W3C ‘Last Call’
  - W3C Recommendation by Autumn
  - WG finished end 2013
- Tools and Best Practice @ ITS Interest Group
  - ITS-XLIFF – Liaison with OASIS TC
  - ITS-RDF: NIF, PROV-O – W3C Community Group
  - ITS-CMIS – contribution to open CMS-TMS
  - Editors and CAT tools

# Questions?