



ITS2.0 Implementation Experience in HTML5 with the Spanish Tax Agency

Pedro L. Díez Orzas

Giuseppe Deriard, Pablo Nieto Caride, Pablo Badía, Consuelo Aldana, Félix Fernández
(Linguaserve)

and

Román Díez González
(Spanish Tax Agency)



1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5 and experience in ITS2.0 annotation:
 - a. Automatic annotation of new ITS2.0 metadata
 - b. Reusing custom tags for ITS2.0 metadata annotation
 - c. Manual ITS2.0 annotation
4. Business case
5. Next steps and beyond 2013

(1) The Spanish Tax Agency

- **Spain: General Indicators 2011**

- Spain is a country regionally structured into 17 autonomous communities and 2 autonomous cities with **5 co-official languages**
- Population : 47,190,493 inhabitants (**12.2 % foreign residents**)

- **Mission of the Spanish Tax Agency**

- Effective application of Spain's tax and custom system
- Management of tax resources on behalf of other public administrations when required by Law or Agreements

- **General taxpayer census**

- Individual taxpayers: 46,509,231
- Companies: 2,674,547
- Other organisations: 2,293,939

- **Total taxpayers: 51,477,717**

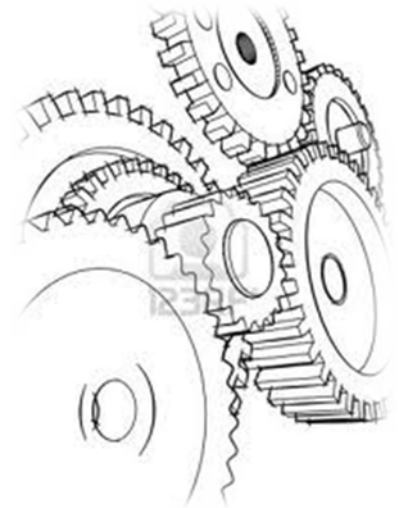




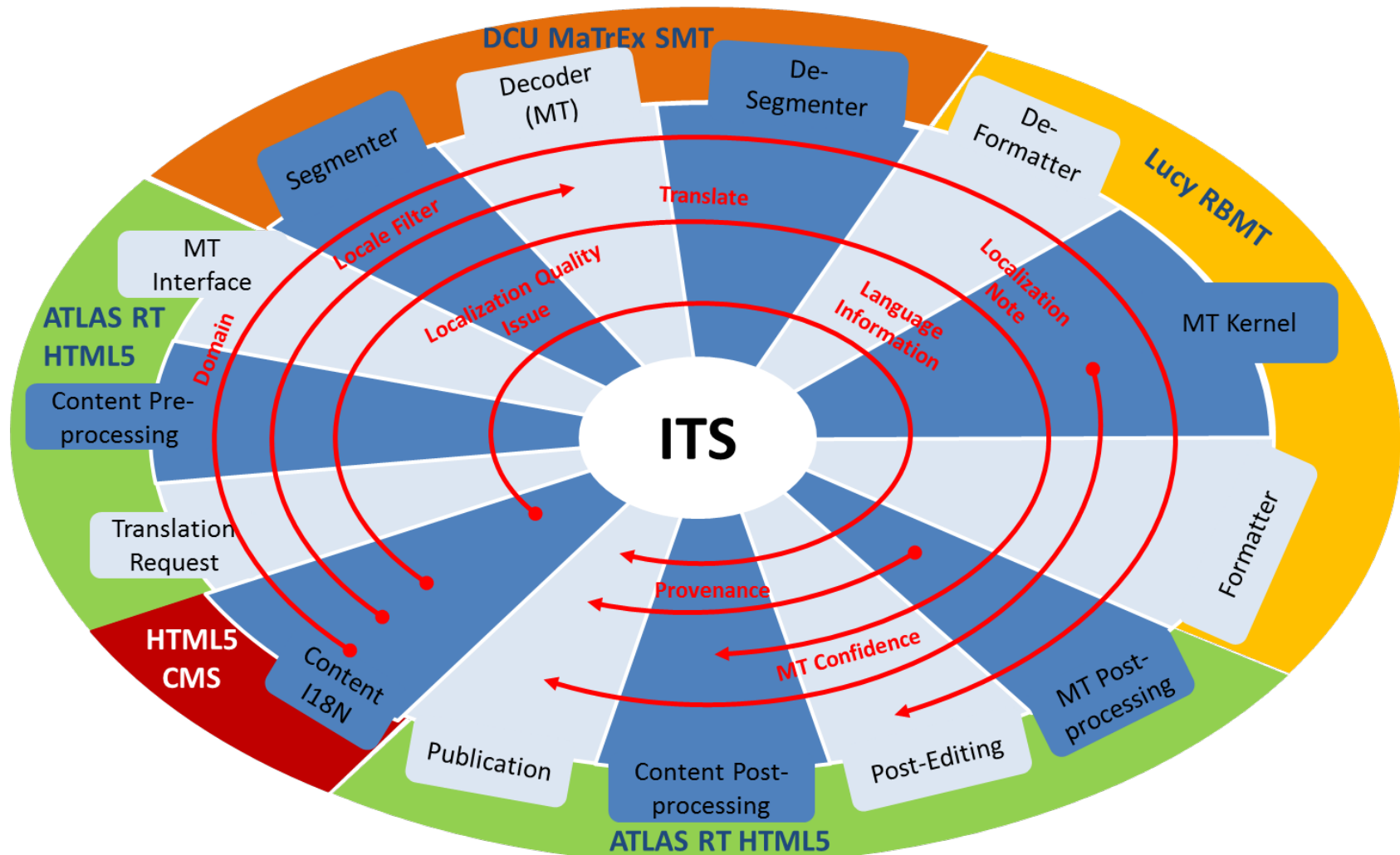
1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5 and experience in ITS2.0 annotation:
 - a. Automatic annotation of new ITS2.0 metadata
 - b. Reusing custom tags for ITS2.0 metadata annotation
 - c. Manual ITS2.0 annotation
4. Business case
5. Next steps and beyond 2013

(2) Spanish Tax Agency in MLW-LT

- **www.agenciatributaria.es**, user in the “Online MT System” use case in the MultilingualWeb-LT (MLW-LT).
- Online MT System use case components:
 - Multilingual www.agenciatributaria.es (CMS: OpenText WEM)
 - HTML5
 - ITS 2.0
 - Real-time Multilingual Publication System
 - ATLAS (Linguaserve’s Real-time Translation System)
 - Lucy Software MT (Rule-based Machine Translation)
 - MaTrEx from Dublin City University (Statistical Machine Translation)



(2) ITS 2.0 in Online MT System I18N



(2) Online MT System I18N

www.agenciatributaria.es



MULTILINGUAL PUBLICATION SYSTEM



Pre-Production/Pre-filters

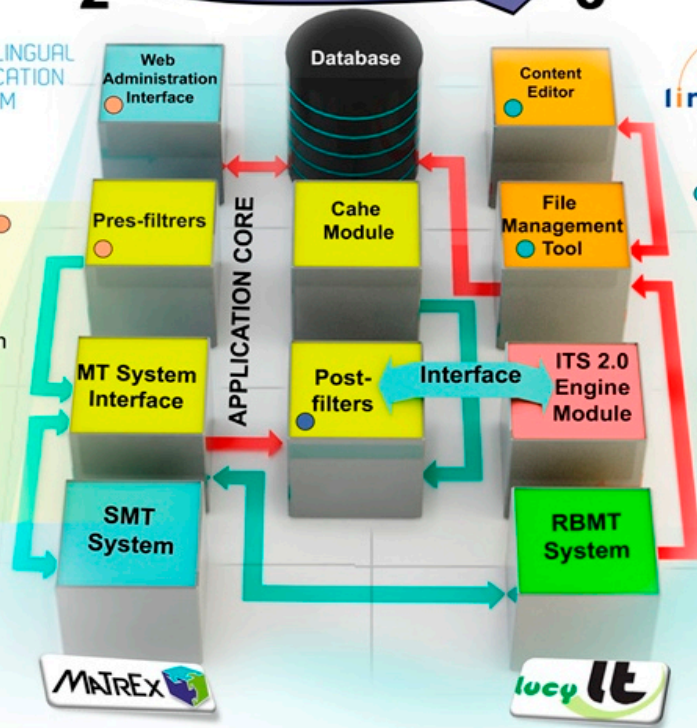
- Translate** ► Adds constant marks to block non-translatable text nodes and attributes.
- Domain** ► Generates a HTML section per domain, and adds such information to the section sent to the MT System to select the appropriate TM and vocabulary (glossary).
- Localization Note** ► Encoding of information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



(2) State of use case (March 2013)

- RTMPS Implementation
 - Prototype 100% (ITS 2.0 definition from Dec 2012)
 - Showcase: pre-production demo
 - ITS 2.0 data categories: 6 (Translate, Localization Note, Language Information, Domain, Provenance, Localization Quality Issue)
- ES-EN total scope: 250 web pages. State:
 - Source language: 30% of target
 - Target language and Post-editing: 30% of target
- ES-FR, ES-DE total scope: 30 web pages. State:
 - Source language: 50% of target
 - Target language and Post-editing: 50% of target
- Testing: pending

(2) Prototypes and Use case

- [ITS 2.0 MaTrEx prototype](#)
- [ITS 2.0 LucySoftware prototype](#)
- [ATLAS Real Time: ITS 2.0 prototype](#)
- [ATLAS Real Time: ITS 2.0 Testing Page](#)
- [Spanish Tax Agency Showcase](#)

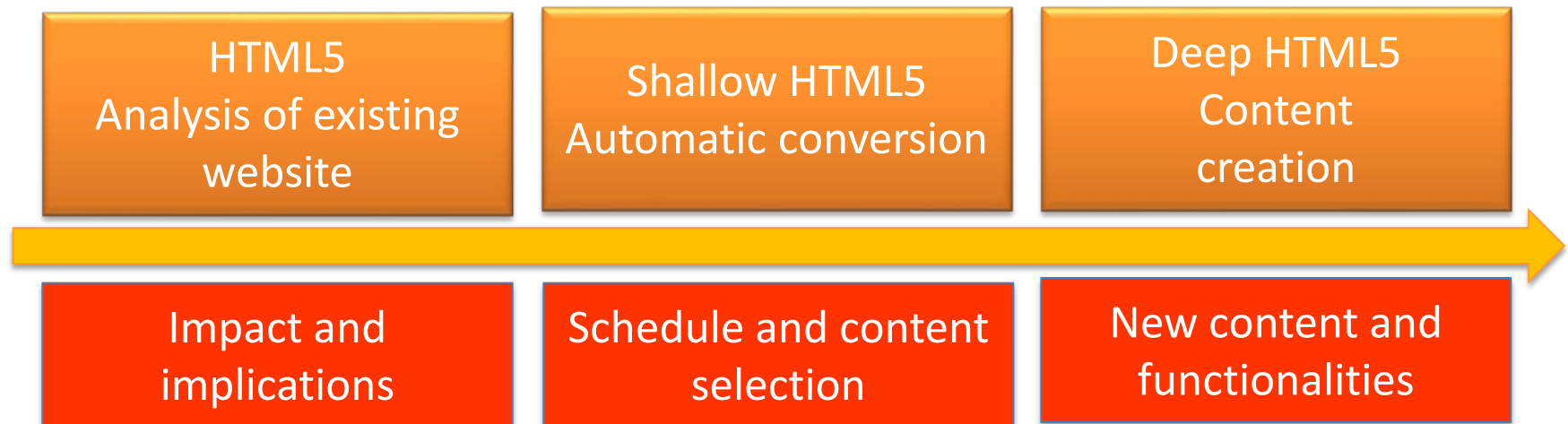




1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5 and experience in ITS2.0 annotation:
 - a. Automatic annotation of new ITS2.0 metadata
 - b. Reusing custom tags for ITS2.0 metadata annotation
 - c. Manual ITS2.0 annotation
4. Business case
5. Next steps and beyond 2013

(3) Shifting to HTML5: Strategy

- Using ITS 2.0 requires HTML version 5 according to the current W3C specification.



(3) Shifting to shallow HTML5: Modifications

- HTML5 DOCTYPE
- The language page (ISO 639-ISO 3166)
- Self-closed tags not allowed
- Head tags
- Erroneous nesting tags
- Attributes separated by spaces
- Non inclusion of presentation attributes in tags
- Header and body structure needed by tables
- HTML entities instead of special characters
- URLs cannot contain special characters
- ID attribute cannot contain spaces
- Required attributes (e.g. tag "object" must always have the attributes "data" and "type")
- Assessed attributes (e.g. "rel" attribute of tags "a" and "link" must be one from a closed list)

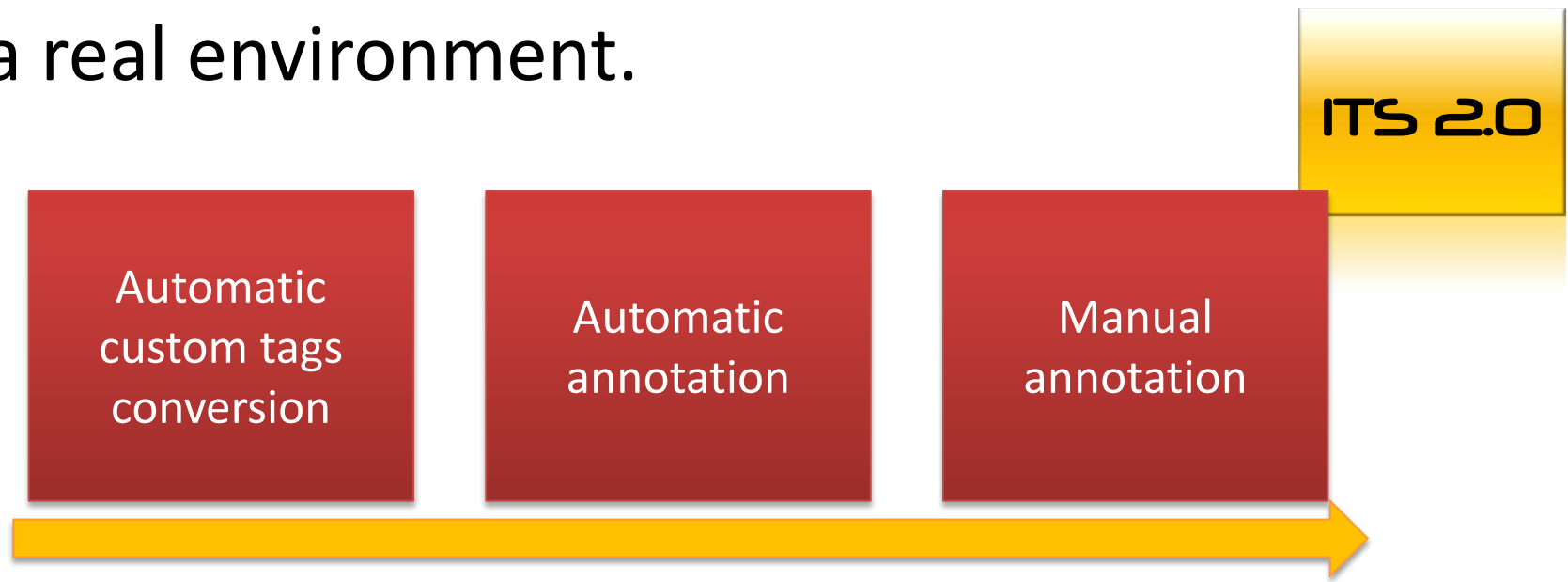


(3) Shifting to shallow HTML5: Obsolete attributes

| Tags | Impact |
|----------------------|--|
| input | Removed the alt attribute from any input tag that does not contain the attribute "type = 'image'" |
| div | Cannot define a "name" attribute in a "DIV" tag |
| a | Not allowed to define the attributes "name" and "title" in tag "a" |
| embed and object | Cannot define the attributes: <ul style="list-style-type: none">•"Applet" in the "embed" and "object" tags•"Name" in the "embed" tag•"Code", "archive", "classid", "codebase", "codetype", "state" and "standby" in the "object" tag |
| table | Not allowed to define the attributes "summary" and "border" in the "table" tag |
| img | Not allowed to define the attributes "name" and "border" in the "img" tag |
| option | Cannot define the attribute "name" in the "option" tag. |
| param | Not allowed to define the attributes "type" and "valuetype" in the "param" tag |
| script | Not allowed to define the attribute "lang" except in "JavaScript", it being case-insensitive in the tag "script" |
| br | Cannot define the attribute "clear" in the "br" tag |
| background attribute | No attribute is used to define the "background" in the tags "body", "table", "thead", "tbody", "tfoot", "tr", "td" and "th". |

(3) ITS2.0 annotation experience

- Strategy adopted in order to annotate the content with ITS2.0 in an efficient and pragmatic way, considering the pressure and requirements of a real environment.



(3) Automatic ITS2.0 reuse of custom tags

- Custom “no translate” tag already exists in the content and is automatically annotated as ITS 2.0 *Translate* data category:

```
<li><!--ATLASP1NOTRAD--><a target="_blank"
href="http://www.boe.es/diario_boe/txt.php?id=BOE-A-2011-20472">Orden EHA/3552/2011,
de 19 de diciembre [...] <!--/ATLASP1NOTRAD--></li>
```



```
<li><a translate="no" target="_blank" href="http://www.boe.es/diario_boe/txt.php?id=BOE-A-
2011-20472">Orden EHA/3552/2011, de 19 de diciembre [...] </li>
```

*Respecting the behaviour of the previous tag and the precedence rules of ITS:

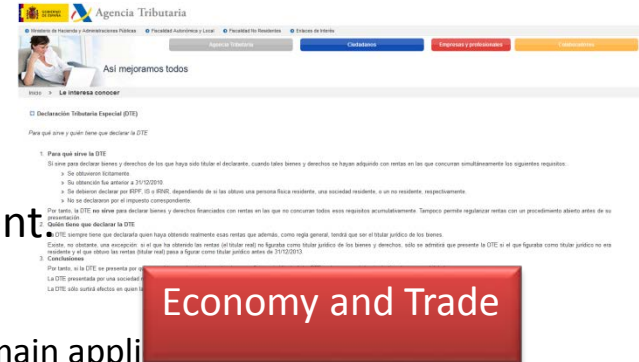
- Addition of ITS default rules for known translatable attributes:
- `<its:translateRule selector="//h:*/@title" translate="yes"/>`
- `<its:translateRule selector="//h:*/@alt" translate="yes"/>`



ITS 2.0

(3) Automatic ITS2.0 annotation: Domain

1. Extracting relevant domains based on the content.
2. Alignment of the domains with each web page.
3. Use of scripts and regular expressions to annotate the content.
4. Document processing:
 - i. The selector points to the html root element, indicating that the domain applies to the whole document (inheritance).
 - ii. The **domainPointer** attribute indicates where the domain that applies to the selected content is ("**Economy and Trade**").
 - iii. The **domainMapping** maps the domain "Economy and Trade" to "**ECON**", which will be sent as an understandable parameter to the MT System.



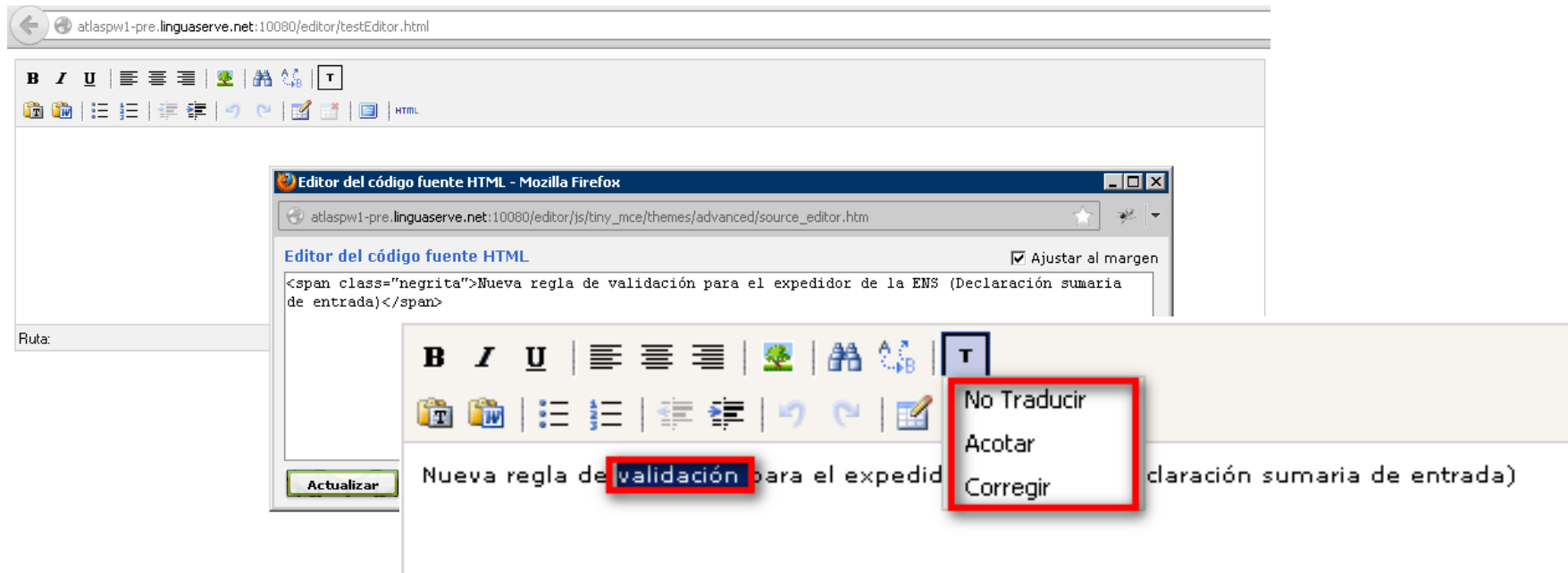
```
<!DOCTYPE html>
<html lang="es">
<head>
<meta charset="utf-8">
<meta name="keywords" content="Economy and Trade"/>
[DOMAIN RULES]
</head>
<body>
[...]
```

```
<its:rules xmlns:its="http://www.w3.org/2005/11/its"
xmlns:h="http://www.w3.org/1999/xhtml" version="2.0">
<its:domainRule
selector="//h:html"
domainPointer="/html/head/meta[@name='keywords']/@content"
domainMapping="Economy and Trade' ECON, 'Law and Legal
Science' LAW, 'General Vocabulary' GV"/>
</its:rules>
```



(3) Manual ITS2.0 annotation

- Quick and pragmatic approach:
 - New HTML Editor plugin created for the ITS 2.0 manual annotation for open source HTML Editor
 - User-friendly interface for the manual insertion of tags.



(3) ITS 2.0 Manual annotation: Translate

- The author must only select the non-translatable element, click on the insertion icon (T) and click on the annotation type: **No Traducir**.

The screenshot shows a web editor interface. At the top, there is a toolbar with various icons, including a 'T' icon for text insertion. A red box highlights the 'No Traducir' option in the dropdown menu that appears when the 'T' icon is clicked. Below the toolbar, the text 'relativas al incumplimiento de dicha limitación.' is visible. Underneath, there is a paragraph of text in Spanish, followed by a list of links: 'Nota informativa', 'Presentar denuncia de pagos en efectivo', and 'Preguntas Frecuentes (INFORMA)'. The 'Preguntas Frecuentes (INFORMA)' link is highlighted with a red box. At the bottom left, there is a 'Volver' link and a breadcrumb trail: 'Ruta: div >> div >> div >> ul.listado >> li >> a >> span'. A dialog box is open in the center, displaying the code 'Insertando INFORMA' and an 'Aceptar' button.



(3) ITS 2.0 Manual annotation: Localization Notes

- Use of the annotation type **Acotar**: The author inserts the annotation text into the box and the software will automatically create the tag.
- The pull-down menu is used to choose the type of localization note. It can either be description (*descriptiva*) or alert (*alerta*).

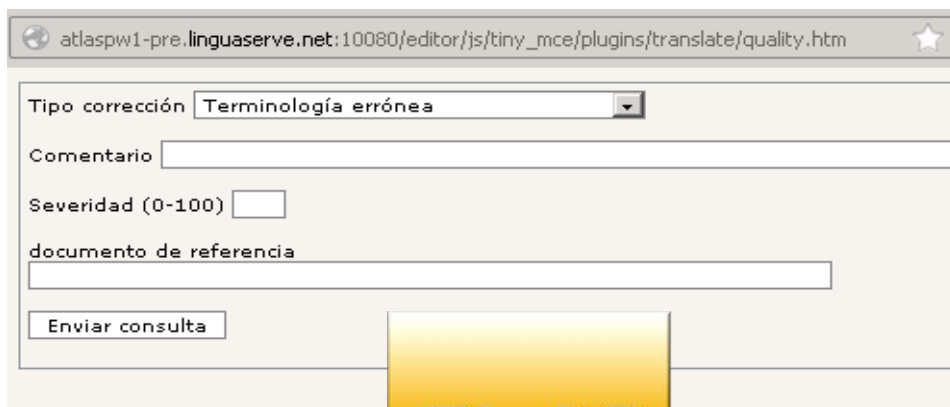
Ruta: ul » li » div » span

<p>La disposición trigésima quinta de la Ley del IRPF

ITS 2.0

(3) ITS 2.0 Manual annotation: Localization Quality Issue

- Use of the annotation type **Corregir**: The author chooses a type of issue from a pull-down menu, inserts a comment into the box (*Comentario*), chooses a severity level between 0 and 100 (*Severidad*) and an optional link to a reference document (*documento de referencia*), and the software will automatically create the tag.



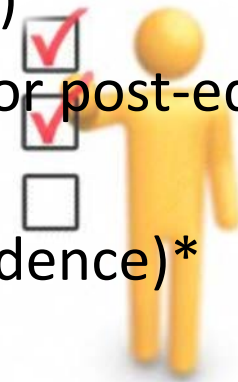
ITS 2.0

Online filing can be done by the interested party or by someone representing them. In both cases, an electronic certificate X.509.V3 issued by the `National Coin and Stamp Factory`

(3) ITS 2.0 benefits

ITS 2.0

- ITS 2.0 Increases user's control and automatic decision processes:
 - Translatability and language pair selection (Translate, Language information)
 - Specific terminology to apply (Domain)
 - Activation rules for post-editing (Localization Note)
 - Quality aspects reported to translation consumer or post-editor (Localization Quality Issue)
 - Post-editors judge quality of translation (MT Confidence)*
 - Identification of agents (provenance)





1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5 and experience in ITS2.0 annotation:
 - a. Automatic annotation of new ITS2.0 metadata
 - b. Reusing custom tags for ITS2.0 metadata annotation
 - c. Manual ITS2.0 annotation
4. Business case
5. Next steps and beyond 2013

(4) MLW-LT Online MT SWOT

Strengths

RTMPS highly reduces:

- Translation costs (Quality on-demand = MT + post-editing %)
- Management costs
- Delivery time
- Technical: clients do not need to install anything

Weaknesses

Viability dependent on :

- Language combination
- MT system output
- Pre-editing and post-editing methodologies and tools (ITS 2.0 and HTML5 compliance)

Opportunities

Profitability:

- Websites with more than half a million words
- Websites with a very high update frequency

Threats

Control, performance and security:

- The client might lose control of the translation
- Real-time performance
- Security level

(4) MLW-LT Online MT Business Case

Strengths

Reduction of:

- Translation costs (MT + post-editing):
 - 100% post-edited: - 30%
 - Depending on % of post-editing cost reduction increases.
- Management costs: -90%
- Non-invasive technology
- Real-time or fast post-edition

Weaknesses

Profitability:

- Sites with more than 4 million words
- Several updates per week

Opportunities

Viability :

- ES>EN (and ES>FR, ES>PT, ES>CA, ES>GL)
- “EDI-TA” methodology and training
- Specific Pre-editing and post-editing tools needed

Threats

Control, performance and security:

- ITS 2.0
- ATLAS RT Cache
- In-house hosting

(4) Business: Opportunities rise from needs

- E-commerce
 - Very high volume and rotation
 - Short texts and repetitive descriptions
 - Better for MT
 - Quicker to post-edit
 - Very sensitive to ITS 2.0 benefits
 - Content source independent (HTML from several CMS and other applications)
- Web 2.0 (user content created)
 - GIST translation
 - Immediacy





1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5 and experience in ITS2.0 annotation:
 - a. Automatic annotation of new ITS2.0 metadata
 - b. Reusing custom tags for ITS2.0 metadata annotation
 - c. Manual ITS2.0 annotation
4. Business case
5. Next steps and beyond 2013

(5) Next steps and beyond 2013

- End of use case (100% of selected scope) – June 2013
 - ES-EN production environment with Lucy LT (under client approval)
 - ES-FR, DE preproduction environment with MaTrEx and Lucy LT
- Exploring best practices using ITS 2.0 data categories.
- Exploring training and methodology in content creation
 - Pre-editing: ITS2.0 capabilities, usage, and training kits.
- Applying training and methodology in translation.
 - EDI-TA : Post-editing: contextual, activation, and identification rules using ITS 2.0.

ITS 2.0

(5) Next steps and beyond 2013

Extensions and implementations

• Readiness

- ITS 2.0 extension data category proposal.
- Linguaserve is applying Readiness in both use cases involved:
 - Applied in CMS-TMS showcase (WP3, poster 3)
 - Applicability in Online Translation system (WP4)
- It indicates the readiness of a document for submission to L10n processes or provides an estimate of when it will be ready for a particular process.
- It can be used in expert systems for automatic processing.

(5) Next steps and beyond 2013

- **Readiness data model**

- *ready-to-process* – type of process to be performed next
- *process-ref* – a pointer to an external set of process type definitions used for ready-to-process
- *ready-at* – defines the time the content is ready for the process, it could be some time in the past, or some time in the future
- *revised* – indicates is this is a different version of the content that was previously marked as ready for the declared process
- *priority* – the priority of the content for the process
- *complete-by* –target date-time attribute for completing the process



(5) Next steps and beyond 2013

Source and target language tools

•Pre-editing:

- Full HTML5 compliance and ITS2.0 annotation facilities
- Writing tools for content quality, and controlled language for post-editing output adaptation

•Post-editing:

- Requirements from EDI-TA and showcase experience
- Specific language-dependent and language-independent post-editing rules and functionalities.
- ITS 2.0 assistance and viewing functions for post-editors.

ITS 2.0

Thank you.

