MultilingualWeb – Language Technology

A New W3C Working Group
Felix Sasaki, David Filip, David Lewis
MultilingualWeb-LT

• New W3C Working Group under I18n Activity  
  – http://www.w3.org/International/multilingualweb/LT/

• **Aims:** define meta-data for web content that facilitates its interaction with language technologies and localization processes.

• Already have 28 participants from 20 organisations  
  – Chairs: Felix Sasaki, David Filip, Dave Lewis

• **Timeline:**
  – Feature Freeze Nov 2012
  – Recommendation complete Dec 2013
Approach

• Standardise Data Categories
  – ITS (1.0) has: Translate, Loc note, Terminology, Directionality, Ruby, Language Info, Element Within Text
  – MLW-LT could add: MT-specific instructions, quality-related provenance, legal?

• Map to formats
  – ITS focussed on XML
    • useful for XHTML, DITA, DocBook
  – MLW-LT also targets HTML5 and CMS-based ‘deep web’
  – Use of microdata and RDFa
Candidate Stakeholders

• Content Author
• CMS-based
  – Localisation Management
  – Translator/Posteditor/Reviewer
• LSP-based (CAT/TMS users)
  – Translator/Posteditor/Reviewer
  – Translation/Review Process Manager
• MT Service Provider
• Text Analytics Service Provider
• CMS Developer
• Localisation Tool developer
• Systems Integrator
• Search engine crawler
• Content Consumer
Scope of Use Cases

Language Technology

Create Content

Translate Content

Consume Content

Language Resources
Source Content Processing

Language Technology
- Named entity recognition

Translation

Create
- Author
- Identify no-translate
- Identify terms
- Localisation Preparation

Language Resources
- Glossary
- Term-base

<..> = Possible MLW-LT Metadata
Localisation Quality Assurance

Create

Translate

Language Technology

Localisation Preparation

Postediting

Translation Review

Translation Memory

Translation Memory+

Language Resources

Term-base

Publish to CMS

Consume Content

Machine Translation

<..> = MLW-LT Metadata
CMS-L10N integration via RDF & XLIFF

Apache Web Server: Servlet container

- RDF Provenance Visualiser
- Sesame Server
- RDFLogger
- Translation Tool
- MT Service
- User Data

Drupal Web CMS

- Sesame Workbench
- RDF Loggers
- Translation tools
Leverage Target Quality Meta-data

- Consume Content
- Leverage Target Quality Meta-data
- Translate
- Language Technology
- MT Training
- Machine Translation
- Publish to CMS
- Reading/Reusing
- Search Indexing
- Language Resources
- Term-base
- Translation Memory+

<..> = MLW-LT Metadata
Rich Meta-data for TM Leverage

The first XLIFF Symposium 2010 is taking place in Limerick, Ireland on 22nd September Test. It will be part of the LRC preconference Test. 1st XLIFF International Test 22 September 2010, Limerick, Test
The 1st International XLIFF Symposium will

<table>
<thead>
<tr>
<th>Text/html; charset=UTF-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010 1st XLIFF Internation Symposium - Limerick, Ireland</td>
</tr>
<tr>
<td>2010 Primer Simposio Internacional sobre XLIFF - Limerick, Irlanda</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Text: en</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first XLIFF Symposium 2010 is taking place in Limerick, Ireland on 22nd September Test. It will be part of the LRC preconference Test. 1st XLIFF International Test 22 September 2010, Limerick, Test</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Translations: es</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first XLIFF Symposium 2010 is taking place in Limerick, Ireland on 22nd September Test. It will be part of the LRC preconference Test. 1st XLIFF International Test 22 September 2010, Limerick, Test</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TM Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Match Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>company-name: irc</td>
</tr>
<tr>
<td>contact-email: <a href="mailto:Lucia.Morado@ul.ie">Lucia.Morado@ul.ie</a></td>
</tr>
<tr>
<td>contact-name: Lucia</td>
</tr>
<tr>
<td>date: 2010-05-01T12:00:00Z</td>
</tr>
<tr>
<td>job-id: 1</td>
</tr>
<tr>
<td>original: Symposium2.htm</td>
</tr>
<tr>
<td>source-language: en</td>
</tr>
<tr>
<td>state: translated</td>
</tr>
<tr>
<td>target-language: es</td>
</tr>
<tr>
<td>tool-company: SDL Trados</td>
</tr>
<tr>
<td>tool-id: Trados</td>
</tr>
</tbody>
</table>

Words: 7 Characters: 56 needs-review-translation
Next Steps

• Contribute to MLW-LT requirements gathering
  – Breakout session Friday
  – Feedback on Requirements
    • New ones? Priorities?
    • http://www.w3.org/International/multilingualweb/lt/wiki/Requirements

• Get involved in WG
  – Participate as W3C members
  – Feedback via public list and WG site
  – Requirements Workshop in Dublin in 11-12 June
  – Implementations

• Where next?– mapping the future of the MLW
  MLW-MultiModal Interaction ....
  MLW-Audio-Visual Content ....
  MLW-JavaScript ....