

XLIFF and ITS: A Secret Marriage



Felix Sasaki (DFKI)
Christian Lieske (SAP AG)

1st International XLIFF Symposium 2010

Felix Sasaki

DFKI

Focus on multilingual (XML, RDF, ...) data on the Web

PhD in Computational linguistics

4 years work as W3C staff in Internationalization, Web Services, Multimedia metadata

XML and Semantic Web geek (yes, it is possible!)

Senior researcher at DFKI (German Research Center for Artificial Intelligence) / Univ. of Applied Sciences Potsdam

Head of German-Austrian W3C Office

Christian Lieske Globalization Services, SAP AG

Knowledge Architect

Content engineering and process automation (including evaluation, prototyping and piloting)

Main field of interest: Internationalization, translation approaches and natural language processing

Contributor to standardization at World Wide Web Consortium (W3C) OASIS and elsewhere

Degree in Computer Science with focus on Natural Language Processing and Artificial Intelligence

What is ITS: The "Internationalization Tag Set"



Specification

<http://www.w3.org/TR/its/>

Best Practice Note

<http://www.w3.org/TR/xml-i18n-bp>

W3C ITS Interest Group

<http://www.w3.org/International/its/ig/>

Internationalization and Localization of XML: Introducing "ITS"



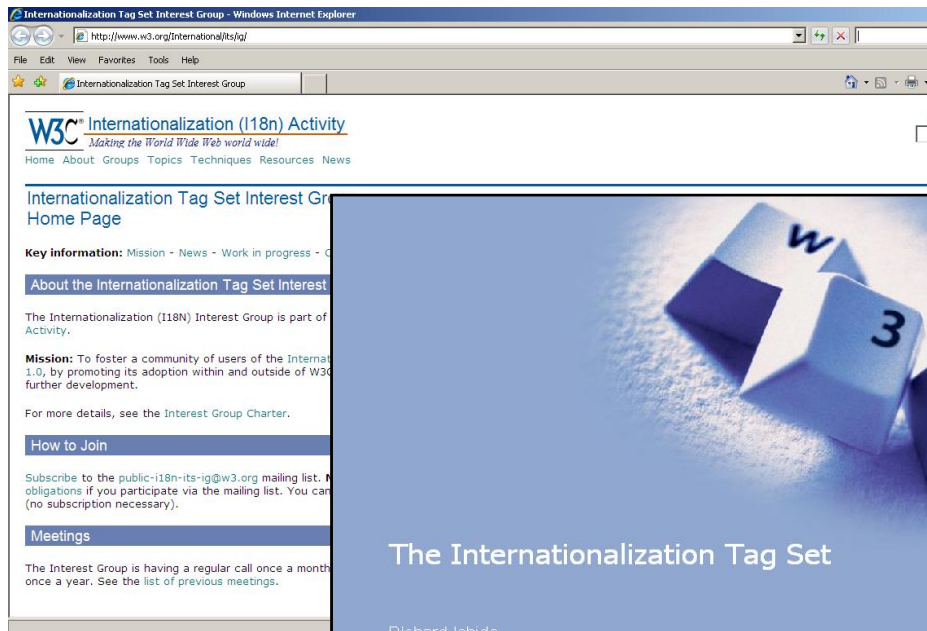
Christian Lieske

Sebastian Rhatz

Felix Sasaki

Slides:

<http://www.w3.org/2006/Talks/0518-xtech-its/>




The Internationalization Tag Set

Richard Ishida
W3C Internationalization Activity Lead

<http://www.w3.org/2006/Talks/10-lrc-its/slides/Slide0010>

Standards-based Translation with W3C ITS and OASIS XLIFF



Christian Lieske (SAP AG)
Felix Sasaki (Fachhochschule Potsdam)
Yves Savourel (Enlaso)
Bryan Schnabel (Tektronix)

tcworld conference 2009

Rhein-Neckar-Hallen Wiesbaden
Thursday, 5th November 2009
8:45 - 10:30 am, Room 1A/3
http://www.tekom.de/upload/2913/LOC12_Sasaki_Lieske.pdf

Fachhochschule Potsdam
University of Applied Sciences
W3C WORLD WIDE WEB
Deutsch-Österr. Büro
THE BEST-RUN BUSINESSES RUN SAP
SAP

Best Practice Note: “Best Practices for XML Internationalization” (not covered here)



This document covers the following requirements:

- [R002 - span-like element](#), see [span](#)
- [R006 - identifying language/locale](#), see [Section 6.7: Language Information](#)
- [R007 - identifying Terms](#), see [Section 6.4: Terminology](#)
- [R008 - purpose specification/mapping](#), see [Section 5.5: Associating ITS Data Categories with Existing Markup](#)
- [R011 - bidirectional text support](#), see [Section 6.5: Directionality](#)
- [R012 - indicator of translatability](#), see [Section 6.2: Translate](#)
- [R014 - limited impact](#), see [Section 5.5: Associating ITS Data Categories with Existing Markup](#)
- [R017 - localization notes](#), see [Section 6.3: Local](#)
- [R020 - annotation markup](#), see [Section 6.6: Rub](#)
- [R025 - elements and segmentation](#), see [Section](#)

The following requirements will be addressed in [\[XML i18n BP\]](#):

- [R003 - CDATA Section](#)
- [R004 - Unique Identifier](#)
- [R005 - Handling of Entities](#)
- [R015 - Attributes and Translatable Text](#)
- [R016 - Naming Scheme](#)
- [R019 - Multilingual Documents](#)
- [R022 - Nested Elements](#)

The Working Group decided not to cover the following

- [R001 - Indicator of Constraints](#)
- [R009 - Content Style](#)
- [R010 - Link to Internal/External Text](#)
- [R013 - Metrics Count](#)
- [R018 - Handling of White-Spaces](#)
- [R021 - Identifying Date and Time](#)
- [R023 - Linguistic Markup](#)
- [R024 - Variables](#)
- [R026 - Associated Objects](#)

important ones.

1. ITS – The “Why?”
2. ITS – The “How?”
3. The big picture: ITS and XLIFF now and in future ...

Who needs support for Internationalization and Localization (of XML)?

- Developers of XML formats
- Process engineers for localization workflows including XML content
- Content producers (including translators) and architects
- Vendors of content-related tools



- Developers: an attribute “translate” for your XML Schema
- Process engineers: tool chain understanding “translate”
- Content producers: people editing or translating content and using “translate”
- Vendors of content-related tools: e.g. CMS producing “translate” with appropriate values for a given format

- Developers: an attribute “dir” for your XML Schema
- Process engineers: tool chain transporting “dir”
- Content producers: people using “dir”
- Vendors of content-related tools: browsers / editors applying “dir” for appropriate visualization

There is a requirement related to I18N / L10N of XML

- “Translate”, “Directionality”, ...

Clear definitions and mechanisms are needed

- What does “Translate”, “Directionality”, ... mean?
- How should it be applied to (XML) content?

Definitions are about metadata

- Often must not disturb original content
- Must be independent of other metadata/data categories

1. ITS – The “Why?”
2. ITS – The “How?”
3. The big picture: ITS and XLIFF now and in future ...

ITS – The “How?” – Objectives



1 Support international use

2 Support localization needs

3 Protect from translatability problems

4 Make meaning of tags easy to recognize

5 Don't disturb

ITS – The “How?” – ITS Mantra



Say important things

- *Do not translate*

About specific content

- *All `uitext` elements*

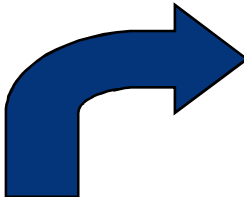
In a standard way

- `its:translate="no"`
- `its:translateRule...`

Mantra Explained (Example: ITS "Translate")



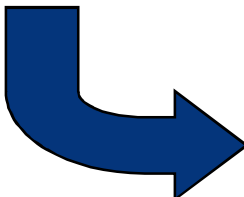
Local Approach



```
<para>  
  Press the  
  <uitext its:translate="no">START</uitext>  
  button to sound the horn. The  
  <uitext its:translate="no">MAKE-READY/ RUN</uitext> indicator flashes.  
</para>
```

```
<para>  
  Press the  
  <uitext>START</uitext>  
  button to sound the horn. The  
  <uitext>MAKE-READY/ RUN</uitext>  
  indicator flashes.  
</para>
```

Global Approach



```
<its:rules ... its:version="1.0">  
  <its:translateRule selector="//uitext" translate="no"/>  
</its:rules>
```

Mantra Explained (ITS Data Categories)



Translate

Mark whether the content of an element or attribute should be translated or not

Localization Note

Communicate notes to localizers about a particular item of content

Terminology

Mark terms and optionally associate them with information, such as definitions

Directionality

Specify the base writing direction of blocks, embeddings and overrides for the Unicode bidirectional algorithm

Ruby

Provide a short annotation of an associated base text, particularly useful for East Asian languages

Language Information

Express the language of a given piece of content

Elements Within Text

Identify how an element behaves relative to its surrounding text, eg. for text segmentation purposes

The ITS data categories are
valuable in themselves

They cannot just be used in the ITS

They are useful also e.g. as RDF /
for other non-XML data

Quiz 1: ITS "Translate" Local or Global



```
<myDoc xmlns:its="http://www.w3.org/2005/11/its" its:version="1.0">
```

```
  <head its:translate="no">
```

```
    <t> Basic Operation</t>
```

```
    <author>Robert Griphook</author>
```

```
    <rev>v13 2007-10-27</rev>
```

```
  </head>
```

```
  <par>To start open <ins>a <b><ref pointer="42"/></b>
```

```
    </ins>. You should observe the flashing of the indicator <as>Indicators are tiny LED's.</as>
```

```
    <n>Bio Charge</n>.
```

```
  </par>
```

```
  <item>
```

```
    <title>Troubleshooting</title>
```

```
    <dev><![CDATA[The <ui>Plug&Restore</ui> is available via the symbol &#229;.]]></dev>
```

```
    <inf>&lt;span class="h1"&gt;Plug&amp;Restore Library&lt;/span&gt;</inf>
```

```
  </item>
```

```
</myDoc>
```

1. Translate the first translatable item.

its:translate='no'

```
<its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0">
  <its:translateRule selector="/myDoc/head" translate="no"/>
</its:rules>
```


Quiz 2: ITS "Localization Note" Global or Local



```
<myDoc xmlns:its="http://www.w3.org/2005/11/its" its:version="1.0">
```

```
  <head>
```

```
    <t>Basic Operation</t>
```

```
    <author>Robert Griphook</author>
```

```
    <rev>v13 2007-10-27</rev>
```

```
  </head>
```

```
  <par>To start open <ins>a <b><ref pointer="42" its:locNote="An icon referring to the printer menu"/></b>
```

```
    </ins>. You should observe the flashing of the indicator <as>Indicators are tiny LED's.</as>
```

```
    <n>Bio Charge</n>.
```

```
  </par>
```

```
  <item>
```

```
    <title>Troubleshooting</title>
```

```
    <dev><![CDATA[The <ui>Plug&Restore</ui> is available via the symbol &#229;.]]></dev>
```

```
    <inf>&lt;span class="h1"&gt;Plug&amp;Restore Library&lt;/span&gt;</inf>
```

```
  </item>
```

```
</myDoc>
```

```
<its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0">
```

```
  <its:locNoteRule locNoteType="description "
```

```
    selector="//ref[@pointer='42']" locNoteRef="EX-devlocnotes-4.html#42" />
```

```
</its:rules>
```

2. Translate the 'a'.

BP22, its:locNote

Quiz 3: ITS "Elements within Text" Global



```
<myDoc xmlns:its="http://www.w3.org/2005/11/its" its:version="1.0">
```

```
<its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0">
```

```
  <its:withinTextRule selector="//as" withinText="nested"/>
```

```
  <its:withinTextRule selector="//n" withinText="yes"/>
```

```
</its:rules>
```

```
  <head>
```

```
    <t>Basic Operation</t><author>Robert Griphook</author><rev>v13 2007-10-27</rev>
```

```
  </head>
```

```
  <par>To start open <ins>a <b><ref pointer="42"/></b>
```

```
  </ins>. You should observe the flashing of the indicator <as>Indicators are tiny LED's.</as>
```

```
  <n>Bio Charge</n>.
```

```
  </par>
```

```
  <item>
```

```
    <title>Troubleshooting</title>
```

```
    <dev><![CDATA[The <ui>Plug&Restore</ui> is available via the symbol &#229;.]]></dev>
```

```
    <inf>&lt;span class="h1"&gt;Plug&amp;Restore Library&lt;/span&gt;</inf>
```

```
  </item>
```

```
</myDoc>
```

3. Run a spell checker on the second sentence.

its:withinTextRule

1. ITS – The “Why?”
2. ITS – The “How?”
3. The big picture: ITS and XLIFF now and in future ...

How do Content and Standards Interact? How can the Future be Shaped?



Implement different requirements with the same mechanisms

- Global / local approach for “Translate”, “Directionality”, ...

Define data categories clearly and independently

- What does “Translate” mean?
- How keep “Directionality” apart from “Ruby” etc.?

Do not enforce changes (“a secret marriage”)

- Do not disturb original content
- Add more & independent metadata

Where Can the Standards already Interact?



ITS2XLIFF tool, see

<http://fabday.fh-potsdam.de/~sasaki/its/>

converts to and from XLIFF based on ITS

Input 1): XML file with ITS “Translate” information

Output: XLIFF file

Input 2): XLIFF file with translated content

Output: original XML file with translated content

ITS2XLIFF generation version 0.6

The W3C Internationalization Tag Set (ITS) defines data categories and their implementation as a set of elements and attributes called the Internationalization Tag Set (ITS). ITS is designed to be used with schemas to support the internationalization and localization of schemas and documents. Please take a look at [link collection](#) for more ITS-related information (amongst others presentations, and references to implementations).

Aside: The W3C ITS Interest Group gives ample opportunities to get involved.

The [XML Localization Interchange File Format \(XLIFF\)](#) gives any documentation or software provider a single interchange file format that can be understood by any localization provider.

This page allows you to generate XLIFF 1.2 from XML files for which W3C ITS rules is available (local markup or global rules). Currently, only the ITS ITS 1.0 [Translate](#) is supported.

The service/functionality behind the page is the [ITS General Decorator](#), an XSLT-based ITS processor which decorates an input document with information for ITS data categories. To be specific, the service works along the lines of the [XLIFF Extraction and Merging](#).

If you want to test the service, you can for example use the samples in the [ITS Test Suite](#).

Feedback is very welcome - please contact [Felix Sasaki](#).

Warning: This is experimental, a stable version is yet to come.

Last Modified: 02/15/2010 21:18:26

For the generation of an XLIFF file, use this form:

XML File

For the integration of translated content into the original XML file, use this form:

XLIFF File for conversion back to XML

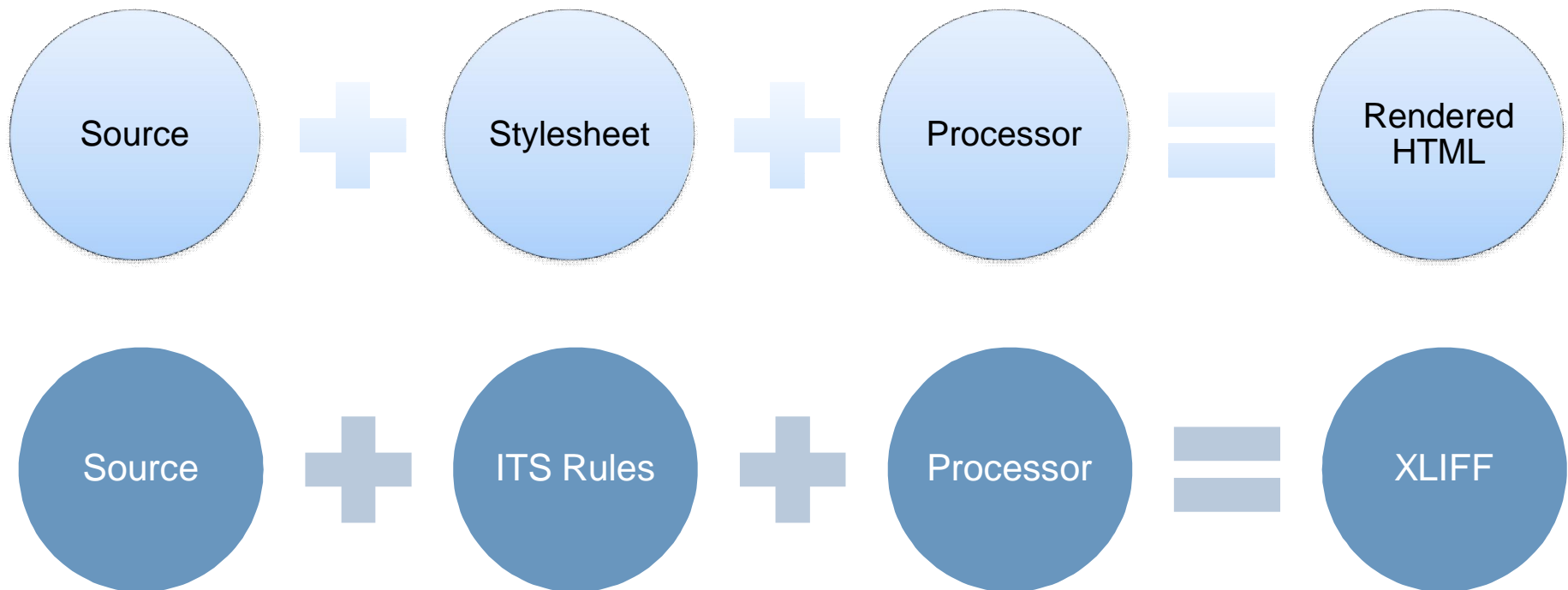
A Vision for the Future of Combining XLIFF and ITS (1/2)



A user agent could use ITS rules for converting content into XLIFF.

Discussion related to a MIME-type for ITS has already been started

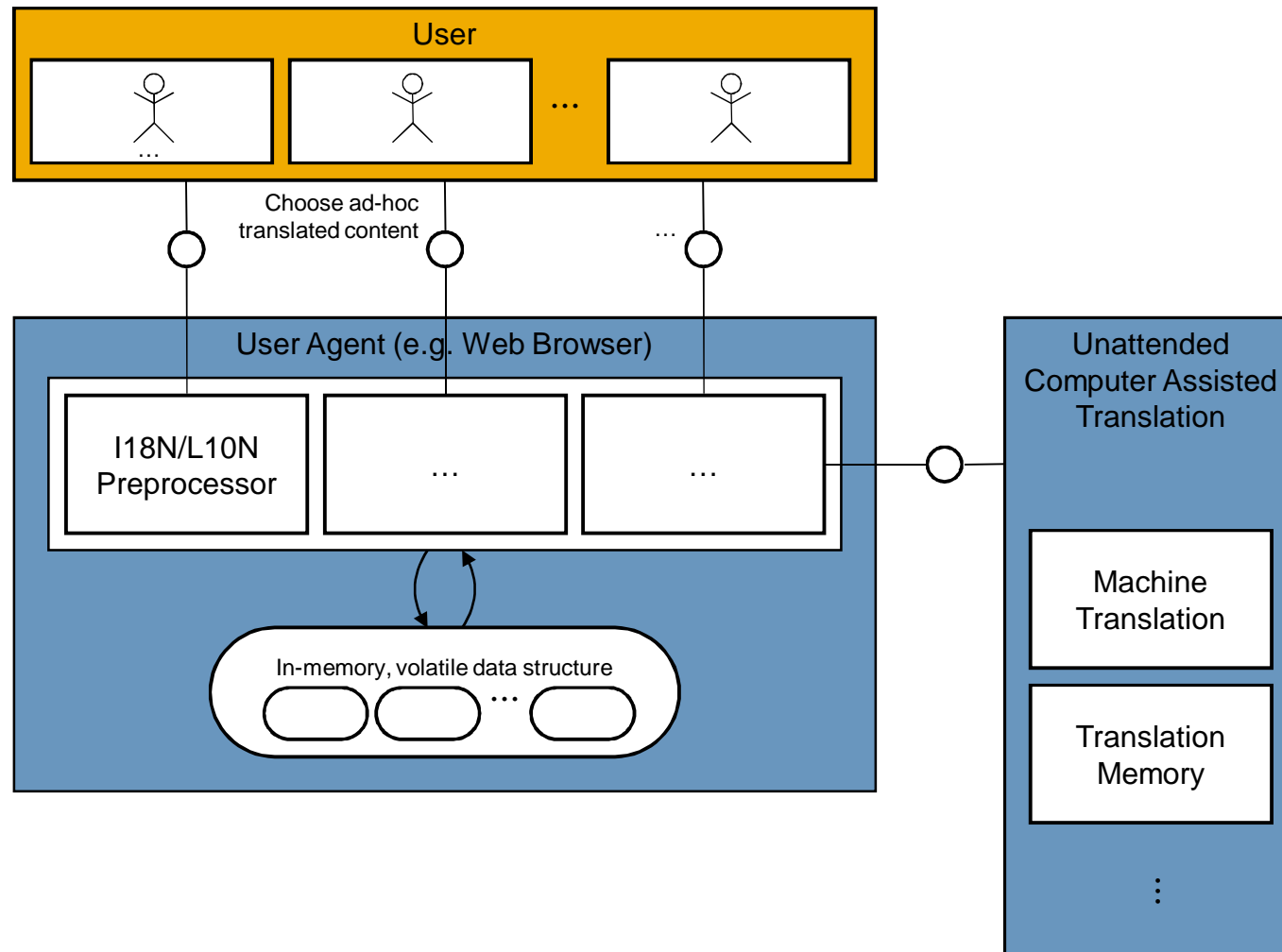
<http://lists.w3.org/Archives/Public/public-i18n-its-ig/2009Jul/0011.html>



A Vision for the Future of Combining XLIFF and ITS (2/2)



Internationalization and Localization for distributed resources based on user clients interpreting ITS and XLIFF



Helping to Shape the Future

The EC funded project “Multilingual Web”



Overall Objective

- Examine how the multilingual Web can be improved through standards and best practices

Dissemination/Outreach

- 4 public events
- Web site (<http://www.multilingualweb.eu>)
- Other items (unfunded, developed by the W3C based on input from partners)
 - Educational material/curriculum
 - Test results (see <http://www.w3.org/International/tests/>)
 - Internationalization Checker (<http://qa-dev.w3.org/i18n-checker>)

Organization

- Thematic Network funded by the European Commission (ICT PSP Grant Agreement No. 250500, and as part of the Competitiveness and Innovation Framework Programme)
- Duration: 24 months from 1 April 2010
- Coordination: World Wide Web Consortium (W3C)/European Research Consortium for Informatics and Mathematics (ERCIM)

Participants

- 22 partners from 15 countries all over Europe
- Industry (providers or users of technology or services), Academia, Standardization Organizations
- Covering wide range of subject areas: language technology, localization, browser development, content creation, social media ...

Dissemination – Public Events



Launch event: *The Multilingual Web –Where are we?*

- 5-6 October 2010
- Madrid, Spain
- <http://www.w3.org/International/multilingualweb/madrid/cfp>

Workshop: *Content Creation/Authoring of the Multilingual Web*

- March 2011
- Pisa, Italy

Workshop: *Translation Tools* (with focus on standards like ITS 1.0, XLIFF, TMX)

- September 2011
- Limerick, Ireland

Workshop: *TBD*

- February 2012
- Luxembourg