W3C Internationalization Tag Set (ITS)

– ITS Widening Doors to Global XML-based Content –

Christian Lieske
SAP AG

Felix Sasaki
W3C

Yves Savourel
ENLASO Corp.
Who will not Present?

Yves Savourel, ENLASO Corporation
- Localization Solutions Architect
- Author of the book *XML Internationalization and Localization*
- In the localization industry for more than 15 years; part of several efforts to take advantage of XML in localization
- One of the architects of XLIFF and TMX
- Chairs the Internationalization Tag Set Working Group at the W3C

Felix Sasaki, World Wide Web Consortium
- Joined the W3C in April 2005
- Works mainly in the Internationalization Activity
- Part of the team at Keio-SFC (Japan)
- Main field of interest is the combined application of W3C technologies for representation and processing of multilingual information
Who will Present?

Christian Lieske, SAP AG

- Natural Language Processing (term extraction and checking, controlled language authoring)
- Content Engineering and Processing (content architecture, application coupling, process design, evaluation, prototyping and piloting)
- W3C (Internationalization Tag Set) and OASIS (XLIFF, Translation Web Services) activities
- Open Lexicon Interchange Format (OLIF)
- Department: Globalization Services
- Degree in Computer Science with focus on Natural Language Processing and Artificial Intelligence
What will be Presented?

Challenges for Global XML Content
How the W3C ITS Helps
ITS and OASIS Source Formats
ITS and Localization

Credits: This presentation uses material from the XLIFF TC, and the W3C ITS WG (notably Richard Ishida)
Challenges for Global XML Content

How the W3C ITS Fills Gaps

ITS and OASIS Source Formats

ITS and Localization
Challenges – Supporting International Use (1/2)

The title says "W3C, פעילות הבינלאומיות" in Hebrew.

<p>The title says "<quote xml:lang="he">פעילה הבינלאומיות, W3C</quote>" in Hebrew.</p>

Result of the Unicode bidi algorithm

The title says "פעילה הבינלאומיות, W3C" in Hebrew.
Volcanic eruptions have literally devastated large inhabited areas. During the 1914 eruption of Sakurajima in Kyushu, 687 houses in Kurokami were buried in hot ash. What remained of this shrine gate, previously five meters tall, was left as a reminder.

Better:

```html
<image src="kk-torii.jpg" height="180" width="240"
caption="Kurokami maibutsu gate (腹五社神社黒神埋没鳥居), Sakurajima Island."
</image>
```
Press the **START** button to sound the horn. The **MAKE READY/RUN** indicator flashes.

Press the **ANLASSEN** button to sound the horn and the **VORBEREITEN** indicator lights up.
<resources>
  <section id="Homepage">
    <arguments>
      <string>page</string>
      <string>childlist</string>
    </arguments>
    <variables>
      <string>POLICY</string>
      <string>Corporate Policy</string>
    </variables>
    <keyvalue_pairs>
      <string>ABC Corp. - Policy Repository</string>
      <string>Footer_Last</string>
      <string>List of Available Policies</string>
    </keyvalue_pairs>
  </section>
</resources>

<dialogue xml:lang="en-gb">
  <rsr id="123">
    <component id="456" type="image">
      <data type="text">images/cancel.gif</data>
      <data type="coordinates">12,20,50,14</data>
    </component>
    <component id="789" type="caption">
      <data type="text">Cancel</data>
      <data type="coordinates">12,34,50,14</data>
    </component>
    <component id="792" type="string">
      <data type="text">Number of files: </data>
    </component>
  </rsr>
</dialogue>
Challenges – The Many Faces of XML (2/2)
### Challenges – Who is Needed to Master them?

<table>
<thead>
<tr>
<th>Role</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema Developers</td>
<td>Create document formats for worldwide use</td>
</tr>
<tr>
<td>Process Engineers</td>
<td>Follow best practices for localization</td>
</tr>
<tr>
<td>Content Producers and Architects</td>
<td>Mark up content for worldwide use</td>
</tr>
<tr>
<td>Vendors Content-Related Tools</td>
<td>Support internationalized formats and content marked up for worldwide use</td>
</tr>
</tbody>
</table>

A standard would make their tasks easier.
The ITS Objectives for Tag Sets

1. **Support international use**
2. **Support localization needs**
3. **Protect from translatability problems**
4. **Make meaning of tags easy to recognize**
The Basic Idea by Example

<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>

<its:rules ... its:version="1.0">
<its:translateRule selector="//uitext" translate="no"/>
</its:rules>
The Basic Idea by Abstraction

Say important things
- Do not translate

About specific content
- All utext elements

In a standard way
- its:translate="no"
### Say Important Things – ITS Data Categories

<table>
<thead>
<tr>
<th>Data Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Translate</strong></td>
<td>Whether the content of an element or attribute should be translated or not</td>
</tr>
<tr>
<td><strong>Localization Note</strong></td>
<td>Communicate notes to localizers about a particular item of content</td>
</tr>
<tr>
<td><strong>Terminology</strong></td>
<td>Mark terms and optionally associate them with information, such as definitions</td>
</tr>
<tr>
<td><strong>Directionality</strong></td>
<td>Specify the base writing direction of blocks, embeddings and overrides for the Unicode bidirectional algorithm</td>
</tr>
<tr>
<td><strong>Ruby</strong></td>
<td>Provide a short annotation of an associated base text, particularly useful for East Asian languages</td>
</tr>
<tr>
<td><strong>Language Information</strong></td>
<td>Express the language of a given piece of content</td>
</tr>
<tr>
<td><strong>Elements Within Text</strong></td>
<td>Identify how an element behaves relative to its surrounding text, eg. for text segmentation purposes</td>
</tr>
</tbody>
</table>
About Specific Content – ITS Selection

Local Approach (Attribute)

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

Global Approach (Element)

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

Can be combined

<its:rules ... its:version="1.0"
xlink:href="myRules-1.xml"/>
For some data categories, special attributes add or point to information about the selected nodes.

```xml
<its:rules its:version="1.0">
  <its:termRule selector="//term" term="yes" termInfoRefPointer="@target"/>
</its:rules>

<p>We may define <term target="#TDPV">discoursal point of view</term> as <gloss xml:id="TDPV">the relationship, expressed through discourse structure, between the implied author or some other addressee, and the fiction.</gloss>
</p>
</text>
```
ITS works with existing schemas or content …

```xml
<its:rules ... its:version="1.0">
  <its:translateRule selector="//*[@change='false']" translate="no/>
  <its:translateRule selector="//*[@change='true']" translate="yes"/>
</its:rules>

<para>
  Press the
  <uitext change="false">START</uitext>
  button to sound the horn. The
  <uitext change="false">MAKE-READY/ RUN</uitext>
  indicator flashes.
</para>
```
Word Wide Web Consortium starts Internationalization Tag Set Working Group

Seminal work by Richard Ishida and Yves Savourel

ITS Specification Working Draft in Last Call stage

ITS Specification transitions to Candidate Recommendation stage

• Provide implementations
• Use the features for localization projects
• Lobby schema designers
Enhancing or Complementing Existing XML-based content

Source/Host Vocabulary
- DITA
- DocBook
- Open Document

Questions
- Benefits
- Division of labour
- Reuse of markup
- Caveats

ITS

W3C Internationalization Tag Set/ Christian Lieske / 22
**Benefits**

### Add missing meta-data
- Directionality
- Ruby

### Formalize knowledge
- Communicate once what needs to be segmented

```xml
<its:rules its:version="1.0">
  <its:withinTextRule withinText="yes" selector="//b | //em | //i"/>
  <its:withinTextRule withinText="nested" selector="//fn"/>
</its:rules>
```

### Enhance granularity
- Indicate that some attributes need to be translated and others not

```xml
<its:rules its:version="1.0">
  <!-- Translatable attribute -->
  <its:translateRule selector="//@alt" translate="yes"/>
</its:rules>
```

### Build on top of ITS processing

![Okapi Framework](image)
Division of Labour

Which constructs from ITS, which from the source/host vocabulary (eg. DITA)?

When you have the choice, use the source/host construct.

Source/host constructs can be associated with ITS data categories.
<topic xmlns:its="http://www.w3.org/2005/11/its" id="myTopic">
<title>The ITS Topic</title>
<prolog>
<its:rules its:version="1.0">
<its:translateRule selector="//[@translate='no']" translate="no"/>
<its:translateRule selector="//[@translate='yes']" translate="yes"/>
<its:termRule selector="/term | //dt"/>
</its:rules>
</prolog>
<!-- body -- >
<body>
<dl>
<dlentry id="tDataCat">
<dt>Data category</dt>
<dd>ITS defines <term>data category</term> as ...</dd>
</dlentry>
</dl>
<p>For the implementation of ITS, ...</p>
<p><ph translate="no" xml:lang="fr">Et voilà !</ph>.</p>
</body>
**Caveats**

**Derivation**

DITA’s specialization, customization, generalization mechanisms have to be compared carefully with the precedence, inheritance, and defaults defined in ITS

**Sample issues related to derivation**

**Inclusion**

Inclusion in DITA is handled by means of the proprietary `conref` mechanism (as opposed eg. to XIInclude)

**Inheritance**

Proliferation rules (e.g. for language information) are defined between DITA maps and other types of DITA objects
Possibilities for DITA and ITS Moving Even Closer to Each Other

- ITS as a DITA Module
- ITS via DITA Specialization
Challenges for Global XML Content
How the W3C ITS Helps
ITS and OASIS Source Formats
ITS and Localization
Globalization/Translation Processes – Actors

Author → Editor → Content Management System → Production Application

Programmer

Translator

Programmer

Disseminator

Product/format X

Product/format Y
Globalization/Translation Processes – Formats

Localizer gets to see …

Native Format
- DITA
- DocBook
- Open Document

Non-native Format
- Proprietary 1
- Proprietary 2
- XLIFF
Globalization/Translation Processes – The Nightmare

For each actor:

For each format:

For each ...

How can you process this (e.g. validation)?

What do you need to know?

What does it look like?
<concept id="myConcept" xml:lang="en-us">
<title>Types of horse</title>
<conbody>
<ol>
<li>Palouse horse:
<p>
<term>Palouse horses</term>
<fn>A palouse horse are same as an <b>Appaloosa</b> have spotted coats. The <term>Nez Perce</term> Indians have been key in breeding this type of horse.</fn>
</p>
</li>
</ol>
</conbody>
</concept>
<its:rules its:version="1.0">
  <!-- Terminology -->
  <its:termRule selector="//term | //dt | //termindex"/>

  <!-- Directionality flags -->
  <its:dirRule selector="//*[dir='ltr']" dir="ltr"/>

  <!-- Elements within text (inline) -->
  <its:withinTextRule withinText="yes" selector="...//boolean | //cite | //itemgroup | //keyword | //ph ...">

  <!-- Non-translatable elements -->
  <translateRule selector="//draft-comment/" translate="no"/>

  <!-- Translatability flags -->
  <translateRule selector="//[@translate='no']" translate="no"/>

  </its:rules>
ITS and Interchange Format – Filter&Merge

Original Material

Filter

Localization Data (Translation Units)

Non-localization Data (Skeleton)

Merge

Translated Material

Localize/Translate

Original Material

Filter

Localization Data (Translation Units)

Non-localization Data (Skeleton)

Merge

Translated Material

Localize/Translate

Example XML code:

```xml
<group restype='menu'>
  <trans-unit id='1' rename='ID_FILEMENU'>
    <source>&amp;File</source>
  </trans-unit>
</group>

<group restype='menuitem'>
  <trans-unit id='1_1' rename='ID_FILE_OPEN'>
    <source>&amp;Open...</source>
  </trans-unit>
  <trans-unit id='1_2' rename='ID_FILE_SAVE'>
    <source>&amp;Save</source>
  </trans-unit>
</group>
```
ITS Test Implementation

This page allows you to test some of the XML Internationalization Tag Set (ITS) rules. See the How to Use This section for information. Last update: Jun-21-2006, 9:02am MDT.

IMPORTANT: This is an early BETA implementation of the Okapi XML Filter and its support for ITS. You may (will?) experience problems with some files.

Enter the XML source (with ITS markup if needed):

```xml
<!-- Example: DITA -->
<concept id="myConcept" xml:lang="en-us">
  <prolog>
    <dita:rules xmlns:dita="http://www.oasis.org/2003/05/tds-implprofile-1.0" dita:version="1.0">
      <dita:translateRule selector="/dita:concept/dita:body/dita:para" translate="yes"/>
    </dita:rules>
  </prolog>
  <dita:conceptTitle title="
    <cobody>
      <p>Palouse horse</p> &amp; palouse horse is the same as an Appalosa.<a href="appalosa.png" alt="Appalosa horse"/><img src="appalosa.png" alt="Appalosa horse"/></a> have spotted coats.</cobody>
  </dita:conceptTitle>
</concept>
```

Extract to XLIFF  Pseudo-Translate  Test Results  Select an example...
Explanations for recording (see previous slide):

- ITS takes the pain out of filtering. You do no longer need to create format-specific filters. Rather, you only need one filter: one that knows ITS. You may even not have to build it yourself, since a free one might be out there.

- Yves Savourel's sample Web page exemplifies, how to easily turn DITA into XLIFF. At its heart: an ITS-aware application which interprets ITS markup in DITA topics.

- Even XLIFF features such as handling of inline markup come through ...
Summary (1/3)

Challenges for Global XML Content

How the W3C ITS Helps

ITS and OASIS Source Formats

ITS and Localization
ITS helps Tag Sets and XML data to ... 

1. Support international use  
2. Support localization needs  
3. Protect from translatability problems  
4. Make meaning of tags easy to recognize
This is your Web – not the W3C's – if something isn't right, get involved to fix it.

Thank you
http://www.w3.org/International/its