XML Internationationalization based on the Best Practice Note of the W3C Internationalization Tag Set Working Group and the W3C Internationalization Tag Set (ITS)

Taking Content on a Safer Global Ride

Christian Lieske (SAP AG)

W3C Day 2010 of the German-Austrian W3C Office
15 September 2010, Berlin

W3C® WORLD WIDE WEB consortium Deutsch-Österr. Büro

Introduction

Taking content global efficiently often fails if best practices are neglected.

In particular, translation processes suffer from lacking internationalization of content, and lacking standardization.

A Working Group (WG) of the World Wide Web Consortium (W3C) addressed this: the Internationalization Tag Set WG. The WG:

- Analyzed needs, and possible pitfalls for global content (in particular XML-related)
- Prioritized
- Devised two important resources for global XML

This presentation sketches background and basics of the two resources

The Recommendation *W3C Internationalization Tag Set (ITS)*

The Working Group note *Best Practices for XML Internationalization (ITS BP)*

Agenda

1. Suffering from Incidents and Accidents

- Failures when content goes global … What's wrong?

2. Performing Root Cause Analysis

- Reasons behind failures … Why is it wrong?

3. Devising Safety Belts

- Research&Development against failures … How approach correction?

4. Driving Safer

- A safe ride without failures ... A better world

Presenter

Christian Lieske

SAP Language Services
Globalization Services
SAP AG

- Knowledge Architect
- Content engineering and process automation (including evaluation, prototyping and piloting)
- Main fields of interest: internationalization, translation approaches and Natural Language Processing
- Contributor to standardization at the World Wide Web Consortium (W3C), OASIS and elsewhere
- Degree in Computer Science with focus on Natural Language Processing and Artificial Intelligence

SAP Software as Backbone of the Global Economy

SAP's solutions run …

Production of **40 million** barrels of oil per day

Retail outlets transactions totaling **$330 million** per day

50 million
Bank accounts with one bank

Production of **32,000** car engines per day

75% of worldwide annual beer production (1.5 billion hectoliter)

54 million
Annual health-care patient visits (US Only)

Defense forces across **107** countries

Processing of **2.5 billion** utility bills per day

65% of worldwide annual chocolate production (**2.2 million** tons)

Production of **4 million** tons of chemicals per day

SAP‘s Standardization Support

SAP founded important standards organizations like WS-I, Eclipse, BIAN and OCEG

SAP plays leading roles in standards organizations like ARTS, HR-XML, The Open Group, OASIS, OAGi, OSGi, JCP, WSI, UN-CEFACT

SAP is an active participant in over 70 industry standards initiatives (messaging, process, collaboration, best practice)

SAP plays a leading role in a number of Open Source projects at Eclipse and Apache

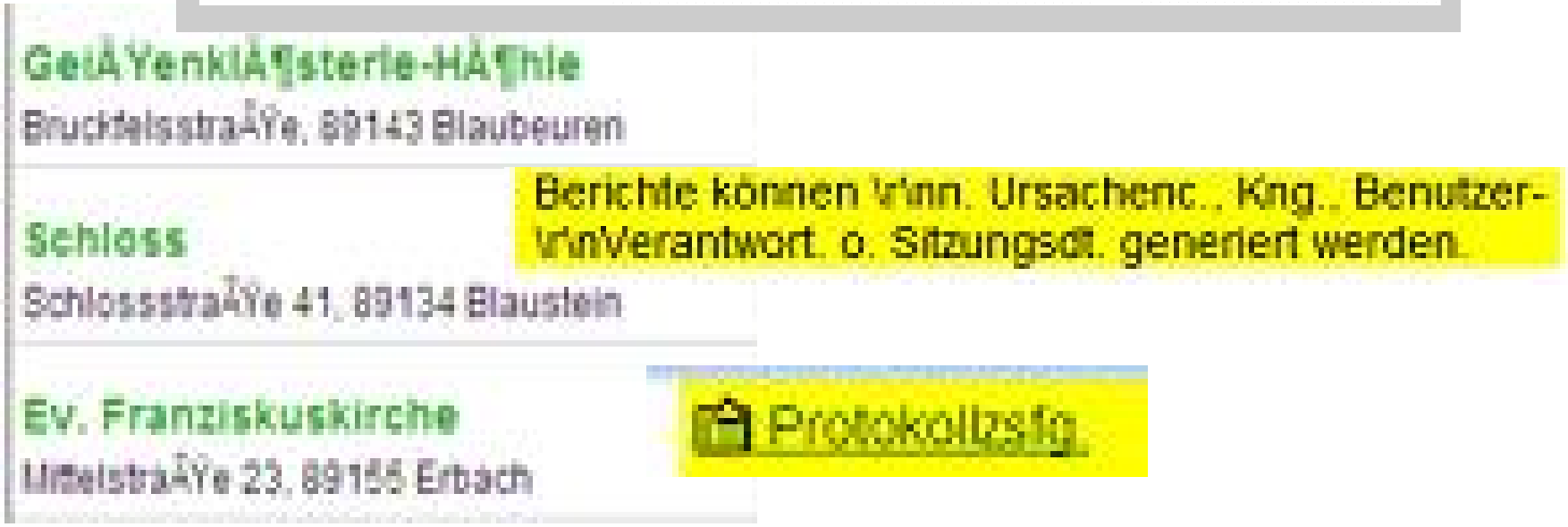
- **Memory Analyzer (MAT)**
- **PAVE**
- **Egit**
- **Aries**

SAP supported Linux back in1999; SAP works closely with most of the key Linux players in the SAP LinuxLab

Suffering from Incidents and Accidents – Example 1

```
<p>
  The title says "
  <quote xml:lang="he">פעילות הבינאום, W3C</quote>
  " in Hebrew.
</p>
```

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



Suffering from Incidents and Accidents – Example 2

```
<myDoc>
        <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>
```

1. Translate the first translatable item.
2. Translate the 'a‘.
3. Run a spell checker on the second sentence.
4. Search for ampersands.

Suffering from Incidents and Accidents – Facets

Translation

Speak the language of the locals

- **Support users by removing language barriers**

Localization

Adapt to culture/environment

- **Legal requirements and statutory reporting**
- **Local best business practices**

Internationalization

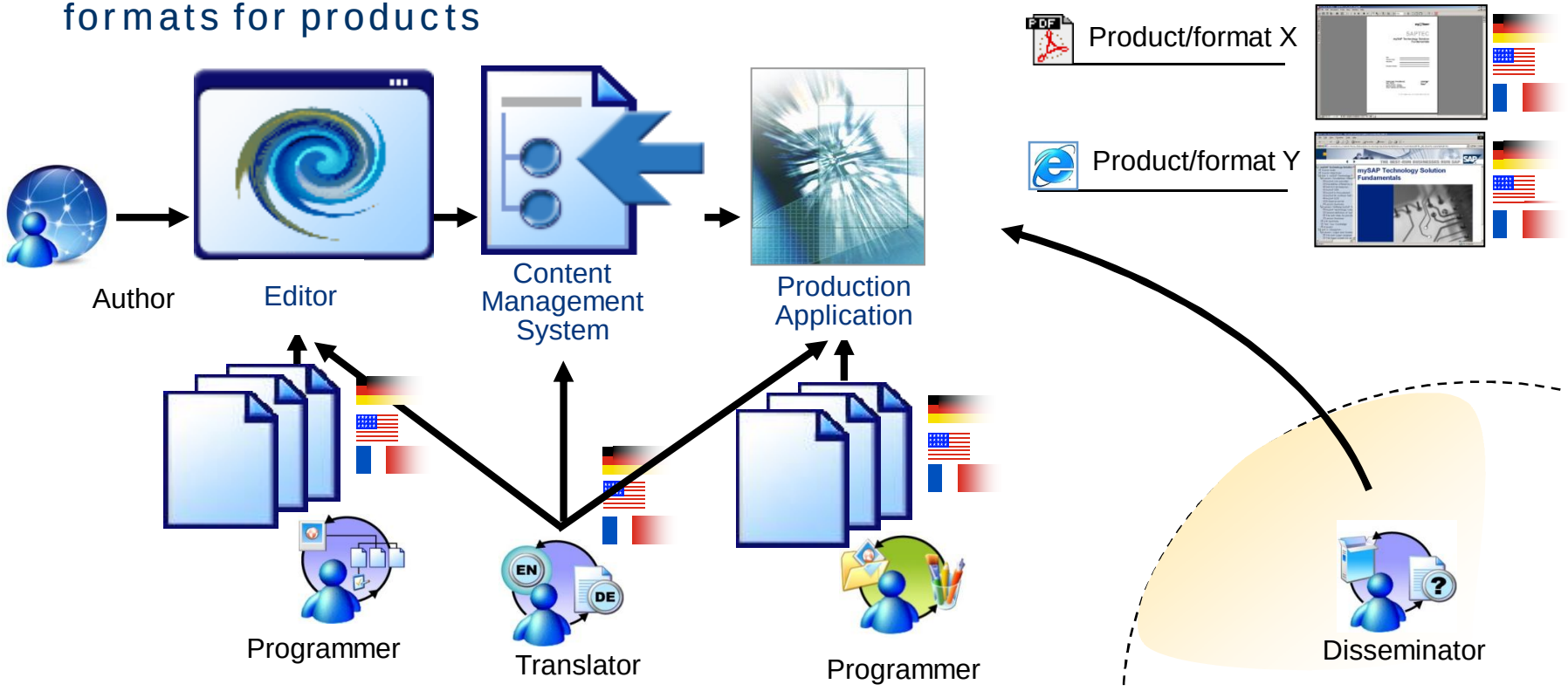
Technical enablement to operate globally

- **Multilanguage support**
- **Fonts, presentation (incl. colours)**
- **Graphics / Images**
- **Code pages / Unicode**
- **Time zones**
- **Multiple currencies**
- **Calendars**

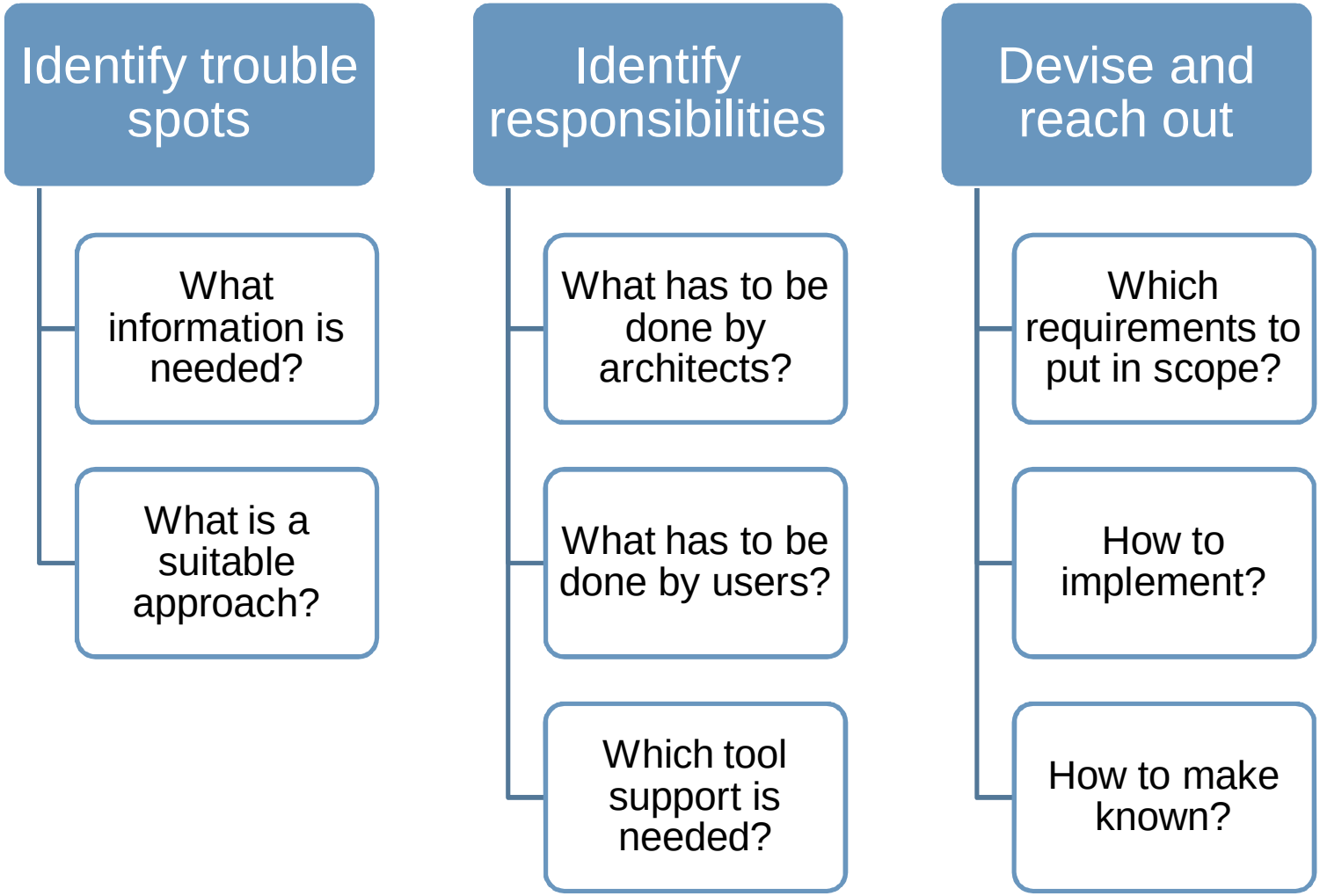
Performing Root Cause Analysis – Areas

Four areas are of particular importance for globalization/translation processes:

1. Source content
2. Collaborative work
3. Coupled applications
4. Languages and formats for products



Devising Safety Belts – Analyze, Prioritize, Devise (1/2)



Performing Root Cause Analysis – Data Failures

Missing data

- What‘s the language?
- Am I allowed to translate this?

Unusable data

- **<t m="n">Basic Operations</>** (unknown semantics)
- **<inf>Plug&Restore Library</inf>** (unusable format)

Wrong data

- ** Hello World! **
- **<script type="text/ecmascript" xlink:href="animate_de.js"/>**
- **<part id="p_de"/>**

Devising Safety Belts – Addressing Data Failures

Meta Data

- Which information do humans or machines need for their work?

Representation and Flow

- How to represent meta data?
- How to move meta data through process?

Evolution

- How to still make sense after 2 years?
- How to deal with versioning, country-variants?

Devising Safety Belts – Analyze, Prioritize, Devise (1/2)

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: Associat
- 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 follo

rtant ones.

- 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

Devising Safety Belts – Analyze, Prioritize, Devise (2/2)

The Recommendation *W3C Internationalization Tag Set (ITS)*

The Working Group note *Best Practices for XML Internationalization (ITS BP)*

Devising Safety Belts – *ITS* 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

Devising Safety Belts – *ITS* Mantra

Say important things

- Do not translate

About specific content

- All *uitext* elements

In a standard way

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

Devising Safety Belts – 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

Devising Safety Belts – ITS Beyond XML

A central dimension of the work done by the ITS Interest Group is related to the general nature of the meta data needs that have been identified.

The ITS data categories are also valuable in other environments (for example when designing meta data for Content Management Systems).

- Translate
- Localization Note
- Terminology
- Directionality
- Ruby
- Language Information
- Elements Within Text

Devising Safety Belts – *ITS* Basic Idea

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>
```

Devising Safety Belts – *Best Practices* Introduction

This document is a complement to the W3C Recommendation *Internationalization Tag Set (ITS) Version 1.0* [ITS]. However, not all internationalization-related issues can be resolved by the special markup described in ITS. The best practices in this document therefore go beyond application of ITS markup to address a number of problems that can be avoided by correctly designing the XML format, and by applying a few additional guidelines when developing content.

When Designing an XML Application

When Authoring XML Content

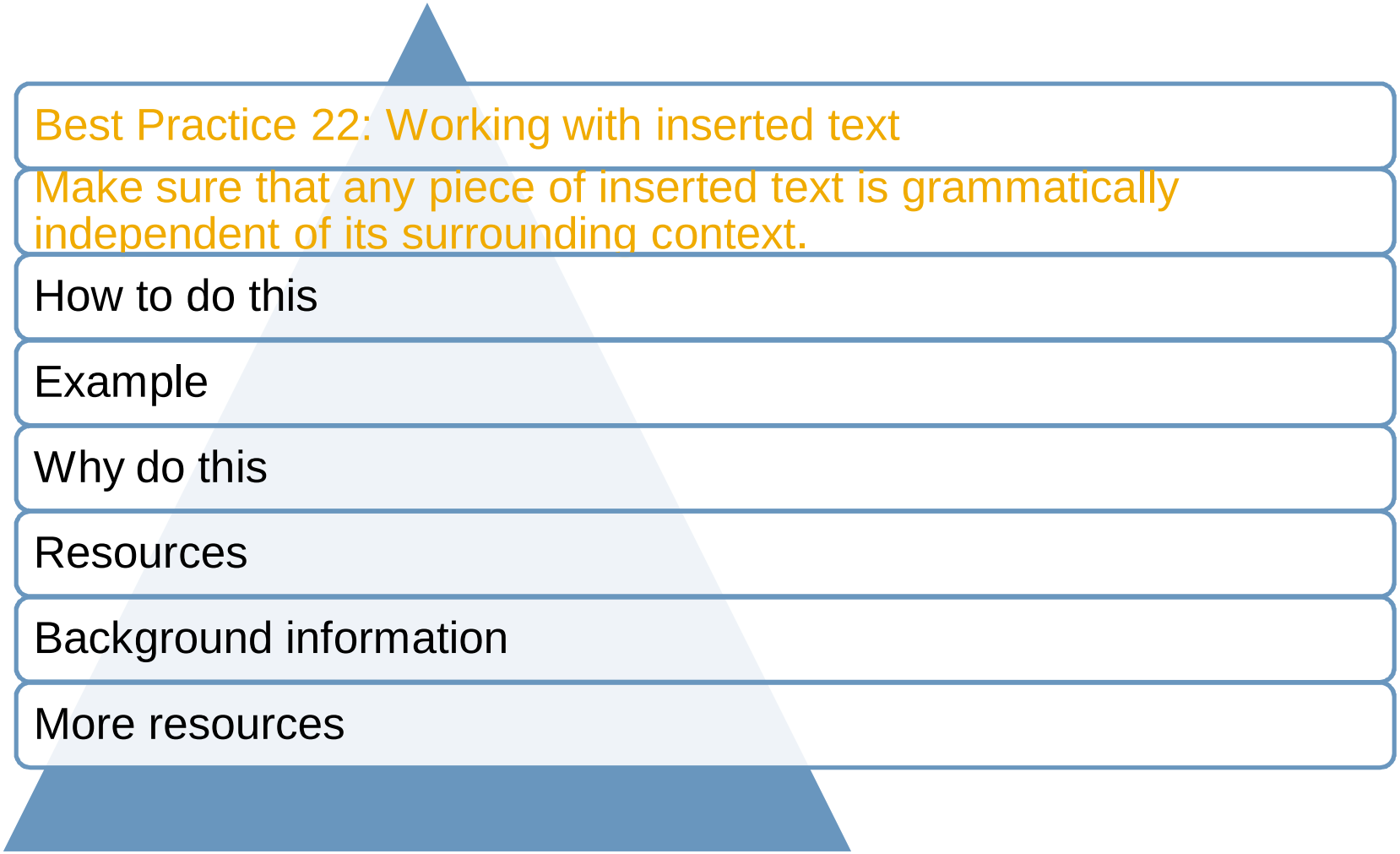
Devising Safety Belts – *BP* Overview (Designing)

- Best Practice 1: Defining markup for natural language labelling
- Best Practice 2: Defining markup to specify text direction
- Best Practice 3: Avoiding translatable attribute values
- Best Practice 4: Indicating which elements and attributes should be translated
- Best Practice 5: Defining markup to override translate information
- Best Practice 6: Providing information related to text segmentation
- Best Practice 7: Defining markup for ruby text
- Best Practice 8: Defining markup for notes to localizers
- Best Practice 9: Defining markup for unique identifiers
- Best Practice 10: Identifying terminology-related elements
- Best Practice 11: Defining markup for specifying or overriding terminology-related information
- Best Practice 12: Working with multilingual documents
- Best Practice 13: Naming elements and attributes
- Best Practice 14: Defining a span-like element
- Best Practice 15: Documenting internationalization and localization features of your schema

Devising Safety Belts – *BP* Overview (Authoring)

- Best Practice 16: Specifying the language of content
- Best Practice 17: Specifying text directionality
- Best Practice 18: Overriding information about what should be translated
- Best Practice 19: Assigning unique identifiers
- Best Practice 20: Avoiding CDATA sections
- Best Practice 21: Providing notes for localizers
- Best Practice 22: Working with inserted text
- Best Practice 23: Identifying terms
- Best Practice 24: Storing markup from another format

Devising Safety Belts – *BP* Internal Structure (1/2)



Devising Safety Belts – *BP* Internal Structure (2/2)

Best Practice 22: Working with inserted text

Make sure that any piece of inserted text is grammatically independent of its surrounding context.

How to do this

Use inserted text only when the text is self-contained and does not affect its surrounding context. For example, titles and quotations are inserted text that, usually, would not cause problems.

Example 28: Providing context to variables.

In this example, in the first message, the element `var` is used to insert the name of a printer. In the second example, it is used to insert a filename. The `its:locNote` attribute is utilized to provide a description of what the variables represent. This may help in deciding how to translate each message.

```
<strings xmlns:its="http://www.w3.org/2005/11/its"
 xml:lang="en" its:version="1.0">
```

Why do this

If not used properly, inserted text can cause important (and sometime

Example 29: Using `conref` in DITA

This is an example of bad design. In this example, the author, working v
the `conref` mechanism. In this case, the term `t123` in `termbase.xm`

```
<p>Using a <term conref="termbase.xml#t123"/>, ra
```

At a first glance the example above seems to work fine in English. Ho

- You should not separate the article from the noun. If "hydraulic lif
article to 'an' or remove it.

Background information

- Internationalization article: Working with Composite Messages. http://www.w3.org/International/articles/composite-messages/
- Internationalization article: Re-using Strings in Scripted Content http://www.w3.org/International/articles/text-reuse/

More resources

Technique index - Topic index

Driving Safer – Example (1/4)

```
<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>
```

Driving Safer – Example (2/4)

```
<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>
```

2. Translate the 'a'.

BP22, its:locNote

```
<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>
```

Driving Safer – Example (3/4)

```
<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

Driving Safer – Example (4/4)

```
<myDoc xmlns:h="http://www.w3.org/1999/xhtml">
        <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&amp;Restore</ui> is available via the symbol &#229;.]]></dev>
                <inf>&lt;span <h:span class="h1„>&gt;Plug&amp;Restore Library&lt;/h:span&gt;></inf>
        </item>
</myDoc>
```

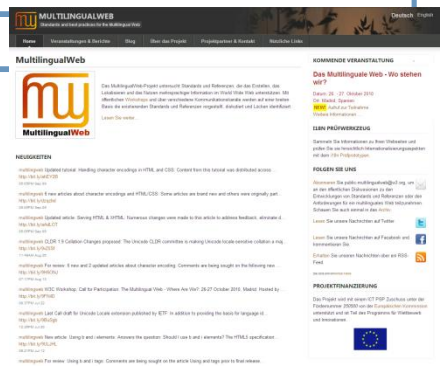
4. Search for ampersands.

   BPs 20, 24

Driving Safer – Outreach (EC Network *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 …

Driving Safer – Outreach (EC Network *Multilingual Web*)

Become involved in ITS and other aspects of the multilingual web

W3C Workshop
The Multilingual Web - Where Are We?
26-27 October 2010, Madrid

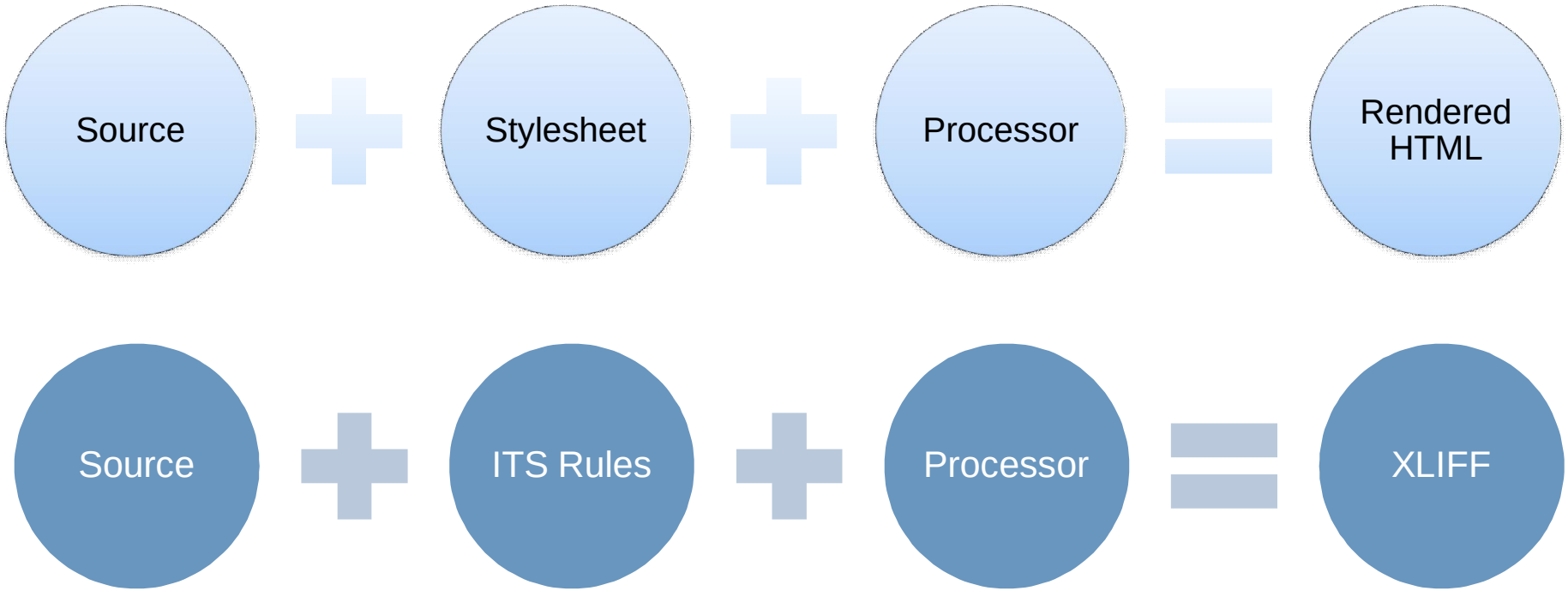
Speaker proposals due 17th September
http://www.w3.org/International/multilingualweb/madrid/cfp

Driving Safer – Outreach (ITS MIME-type)

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

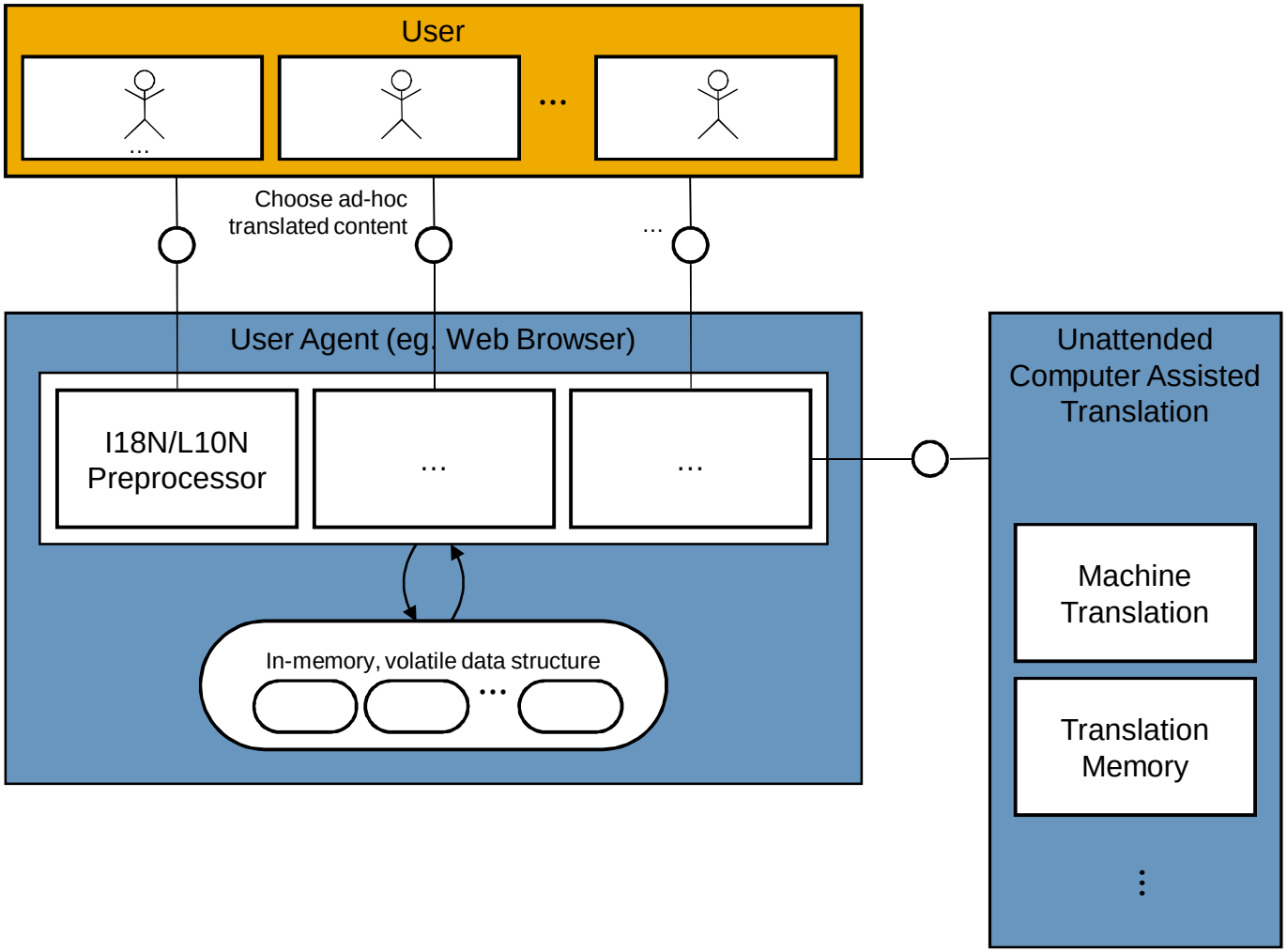
Discussion related to a MIME-type for ITS has already been started

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



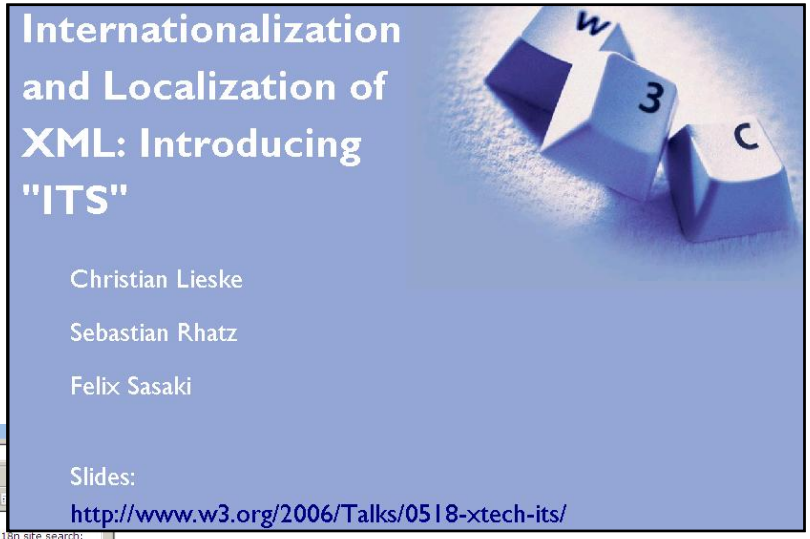
Driving Safer – Outreach (ITS Visions)

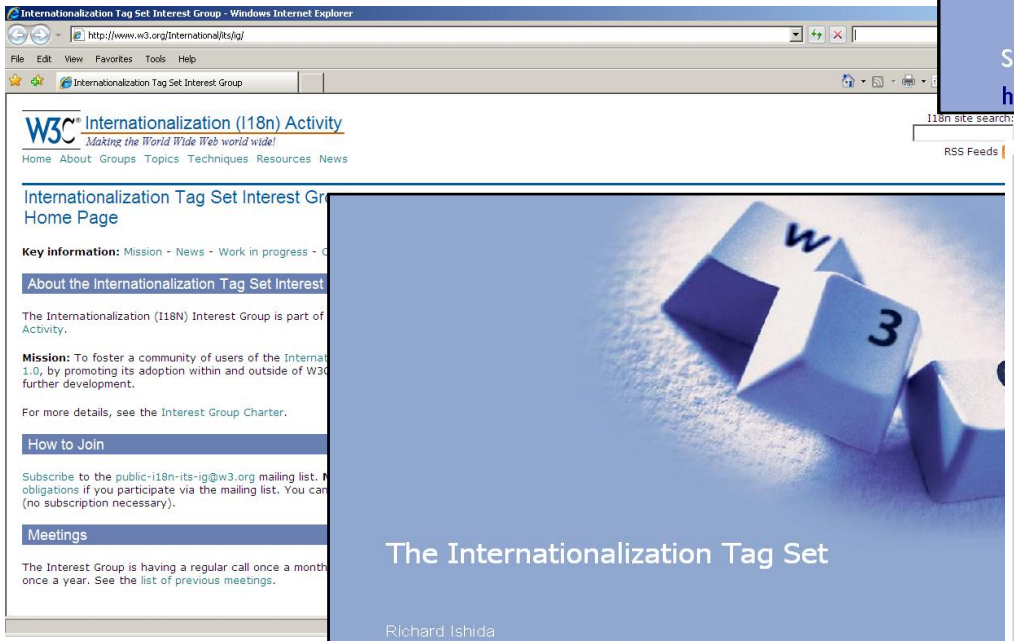
Internationalization and Localization for distributed resources based on user clients interpreting ITS and the XML Localization Interchange File Format (XLIFF).

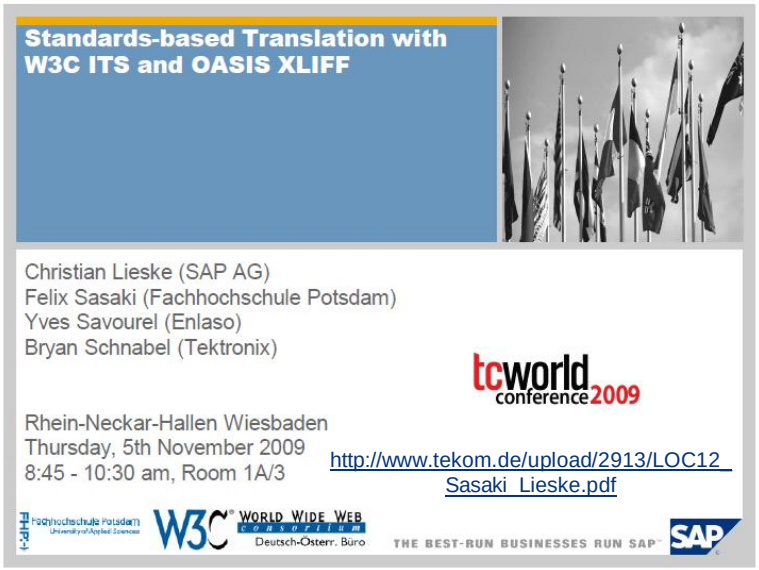


W3C ITS and Best Practices – Further Information

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/







Thank you!

Christian Lieske christian.lieske@sap.com

Disclaimer

All product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

This document may contain only intended strategies, developments, and is not intended to be binding upon the authors or their employers to any particular course of business, product strategy, and/or development. The authors or their employers assume no responsibility for errors or omissions in this document. The authors or their employers do not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

The authors or their employers shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The authors have no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.