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

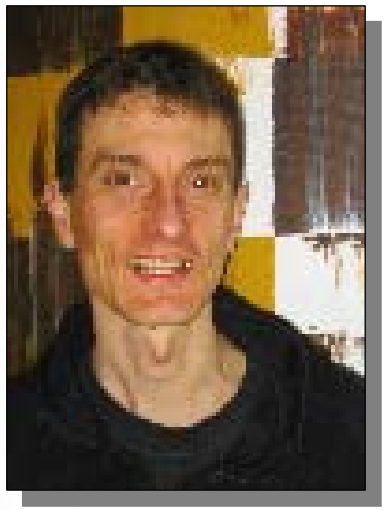
- Localization Solutions Architect
- Author of the book *XML Internationalization and Localization*
- In the localization industry for more than 15 years; part of several efforts to take advantage of XML in localization
- One of the architects of *XLIFF* and *TMX*
- Chairs the Internationalization Tag 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

ENLASO enterprise language solutions

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)

Challenges – Supporting International Use (2/2)

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.

Kurokami maibutsu gate (腹五社神社黒神埋没鳥居), *Sakurajima Island.*

<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." />

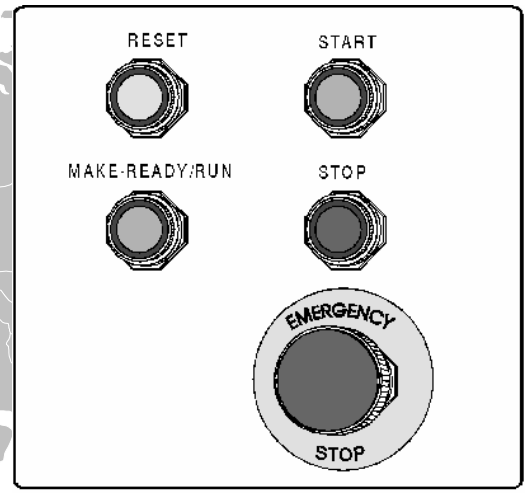
Better:

```
<image src="kk-torii.jpg" height="180" width="240">
        <caption>Kurokami maibutsu gate
        (<span xml:lang="ja">腹五社神社黒神埋没鳥居</span>), Sakurajima Island.</caption>
</image>
```

ENLASO enterprise language solutions

Challenges – Supporting Localization

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



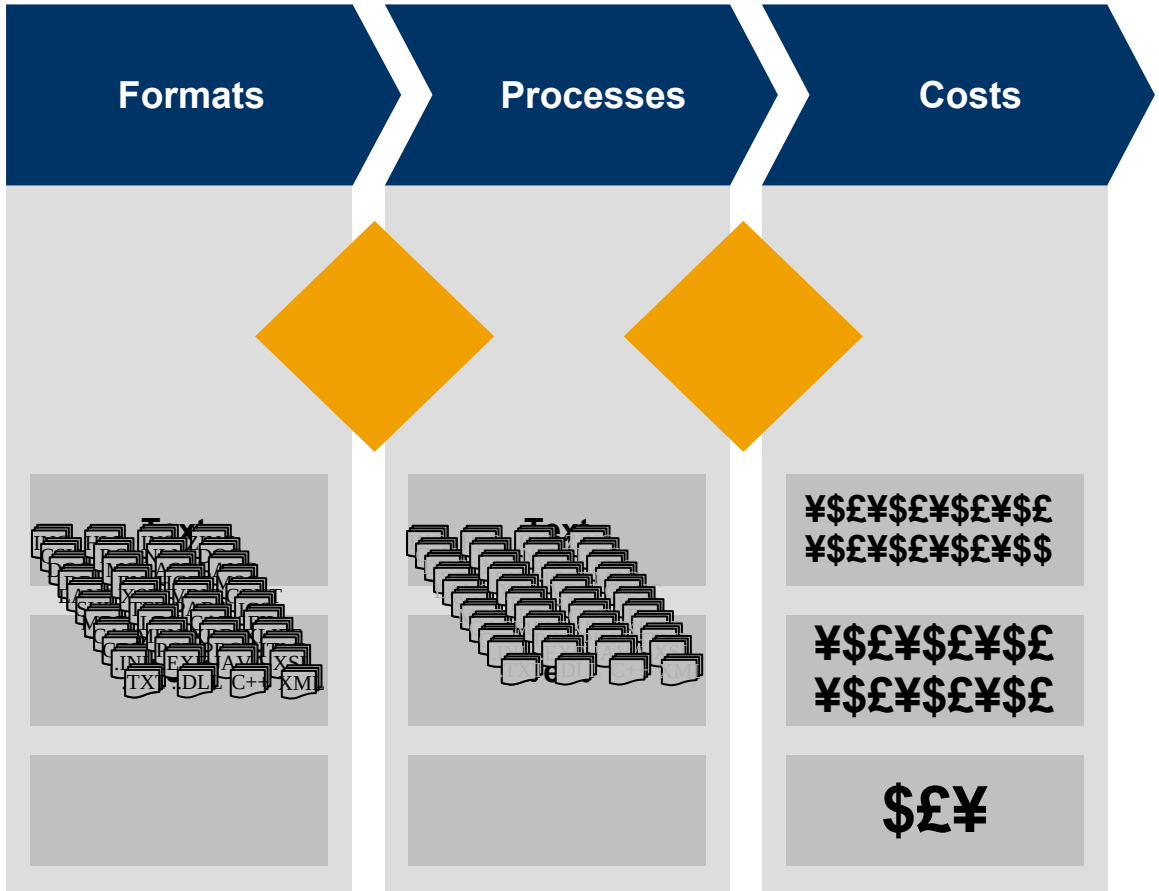
<para>
<uitext>ANLASSEN</uitext>
drücken, so dass die Hupe ertönt und die Anzeige
<uitext>VORBEREITEN</uitext> blinkt.
</para>

Challenges – The Many Faces of XML (1/2)

```
<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">
 <rsrc 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>
 </rsrc>
</dialogue>
```

Challenges – The Many Faces of XML (2/2)



Challenges – Who is Needed to Master them?

Schema Developers	Create document formats for worldwide use
Process Engineers	Follow best practices for localization
Content Producers and Architects	Mark up conte dwide use
Vendors Conten	upport internationalized formats and content marked up for worldwide use

A standard would make their tasks easier

Challenges for Global XML Content

How the W3C ITS Helps

ITS and OASIS Source Formats

ITS and Localization

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

Local Approach

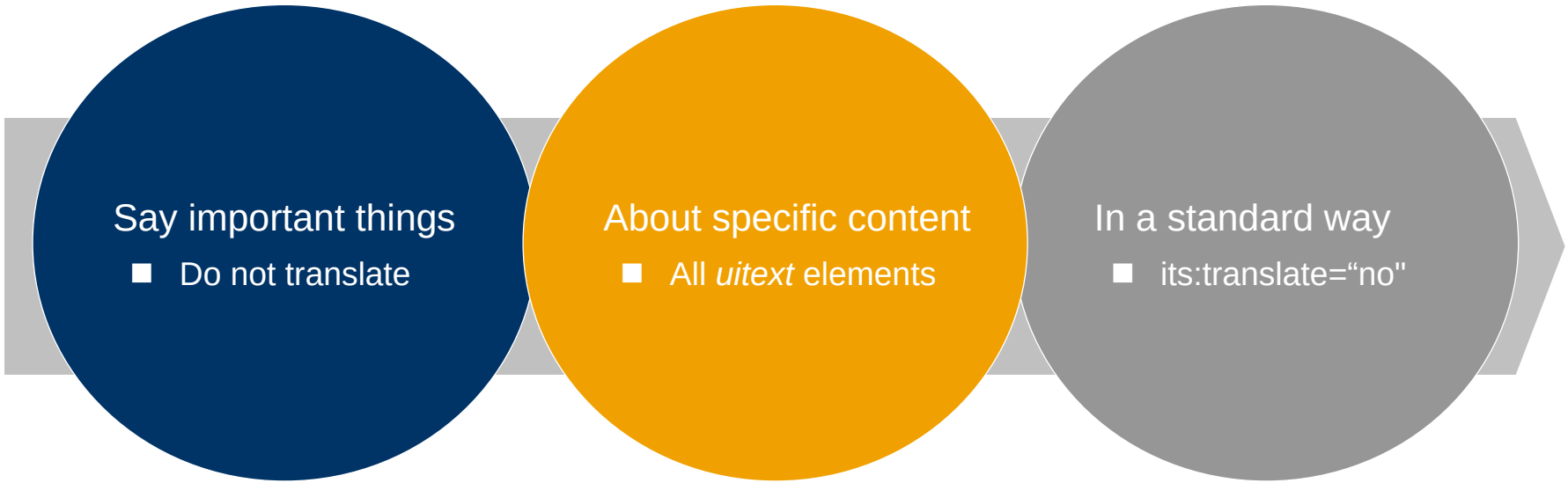
```
<para>
    Press the
    <uitext its:translate="no">START</uitext>
    button to sound the horn. The
    <uitext its:translate="no">MAKE-READY/ RUN</uitext>
    indicator flashes.
</para>
```

```
<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>
```

The Basic Idea by Abstraction



Say Important Things – ITS Data Categories

Data Category

Translate	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

About Specific Content – ITS Selection

Local Approach (Attribute)

```
<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>
```

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"/>
```

In a Standard Way – Adding or Pointing to Information

For some data categories, special attributes add or point to information about the selected nodes.

```
<text>

    <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 addresser, and the
        fiction.</gloss>
    </p>
</text>
```

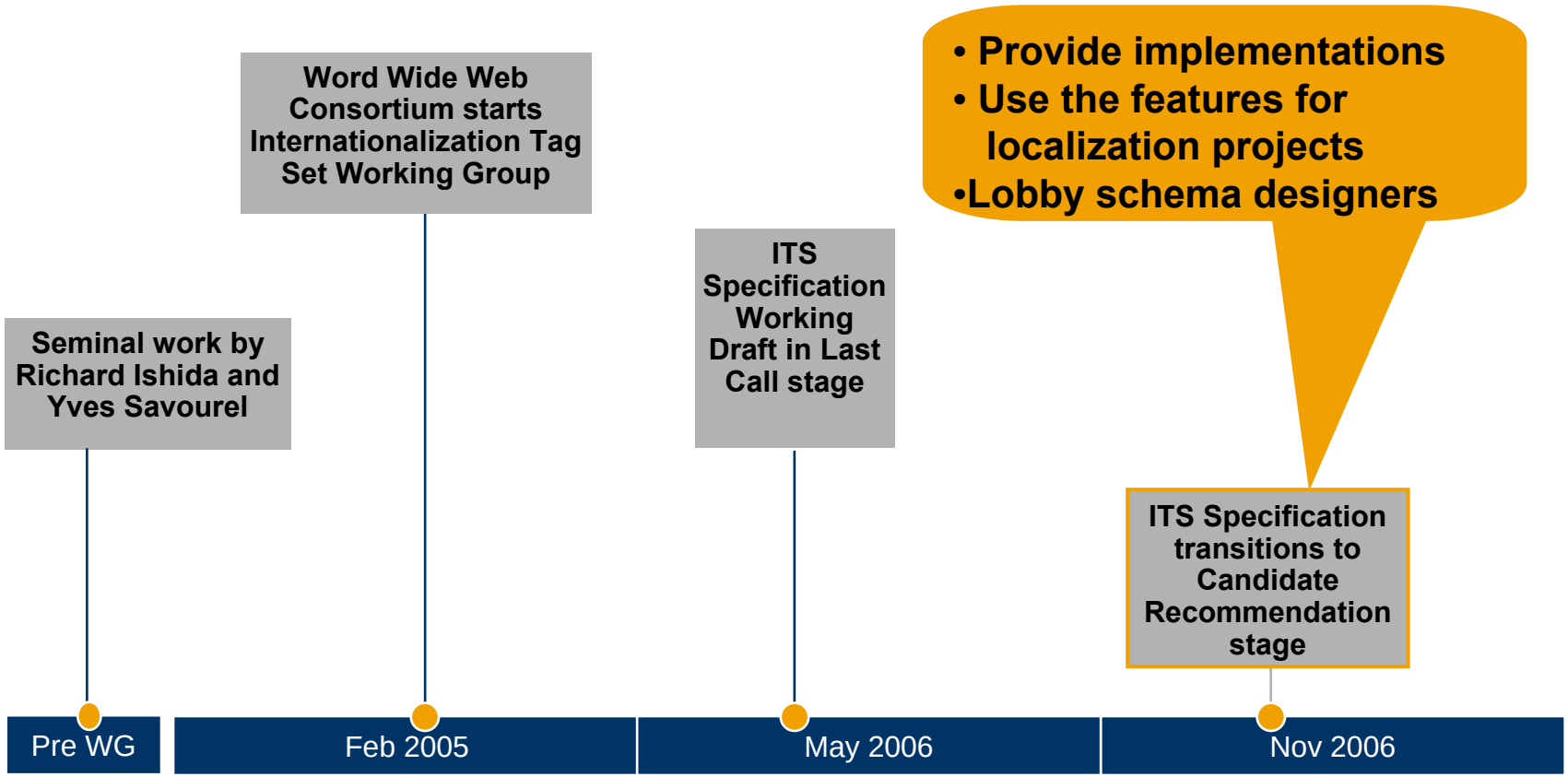
In a Standard Way – Reusing Markup

ITS works with existing schemas or content …

```
<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>
```

In a Standard Way – The Road Travelled and Ahead



Challenges for Global XML Content

How the W3C ITS Helps

ITS and OASIS Source Formats

ITS and Localization

Enhancing or Complementing Existing XML-based content

Source/Host Vocabulary

- DITA
- DocBook
- Open Document

Questions

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

ITS

Benefits

Add missing meta-data

- Directionality
- Ruby

Formalize knowledge

- Communicate once what needs to segmented

```
<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

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

Build on top of ITS processing

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

Reuse of Markup

```
<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 -- >
</topic>
```

```
<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

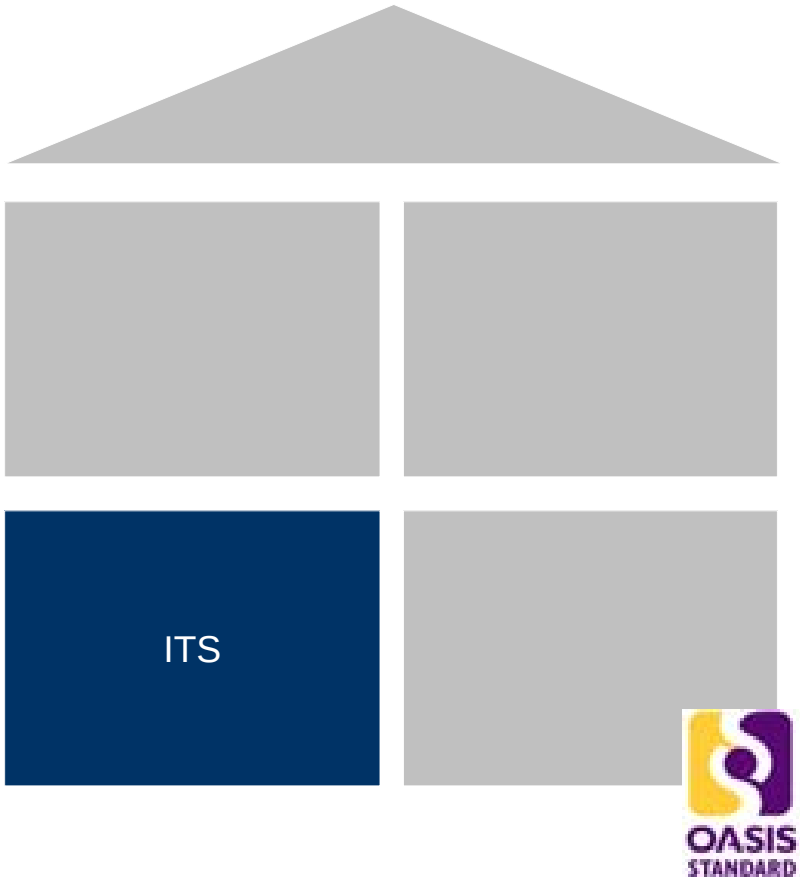
Inclusion

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

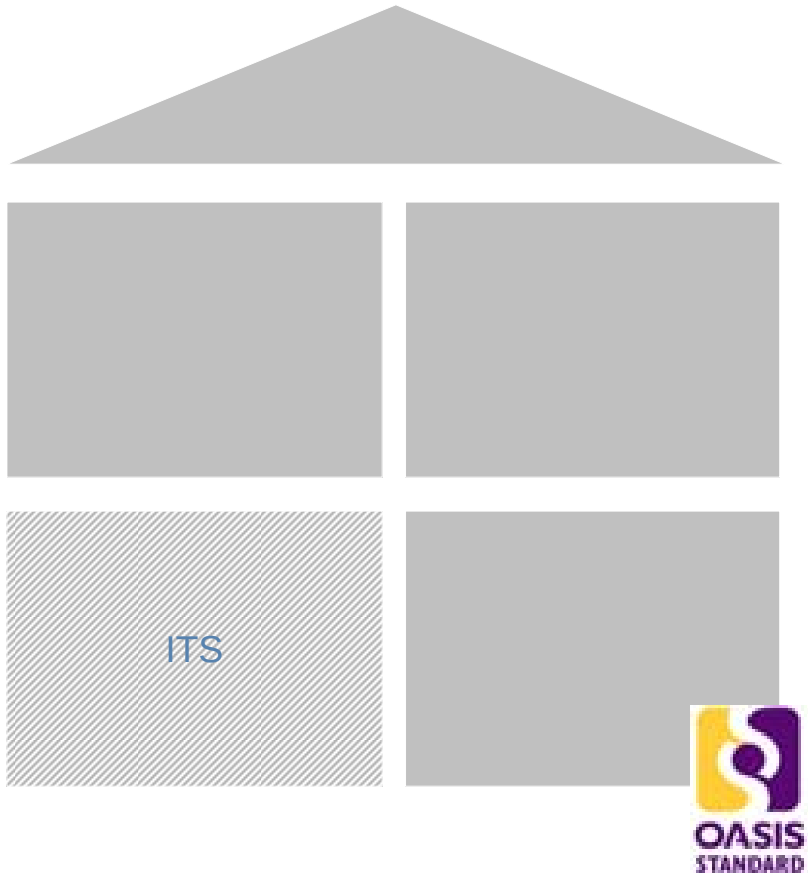
Inheritance

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

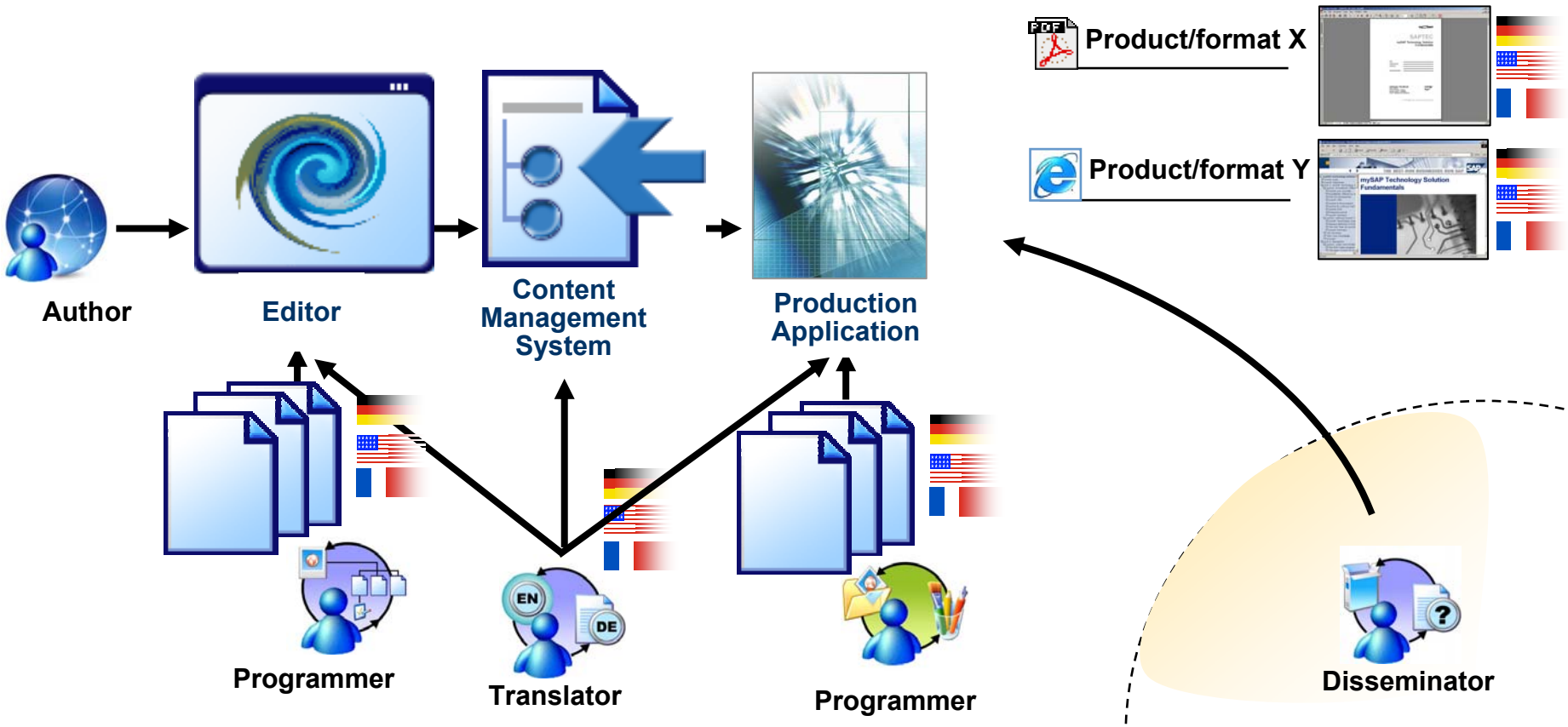
Challenges for Global XML Content

How the W3C ITS Helps

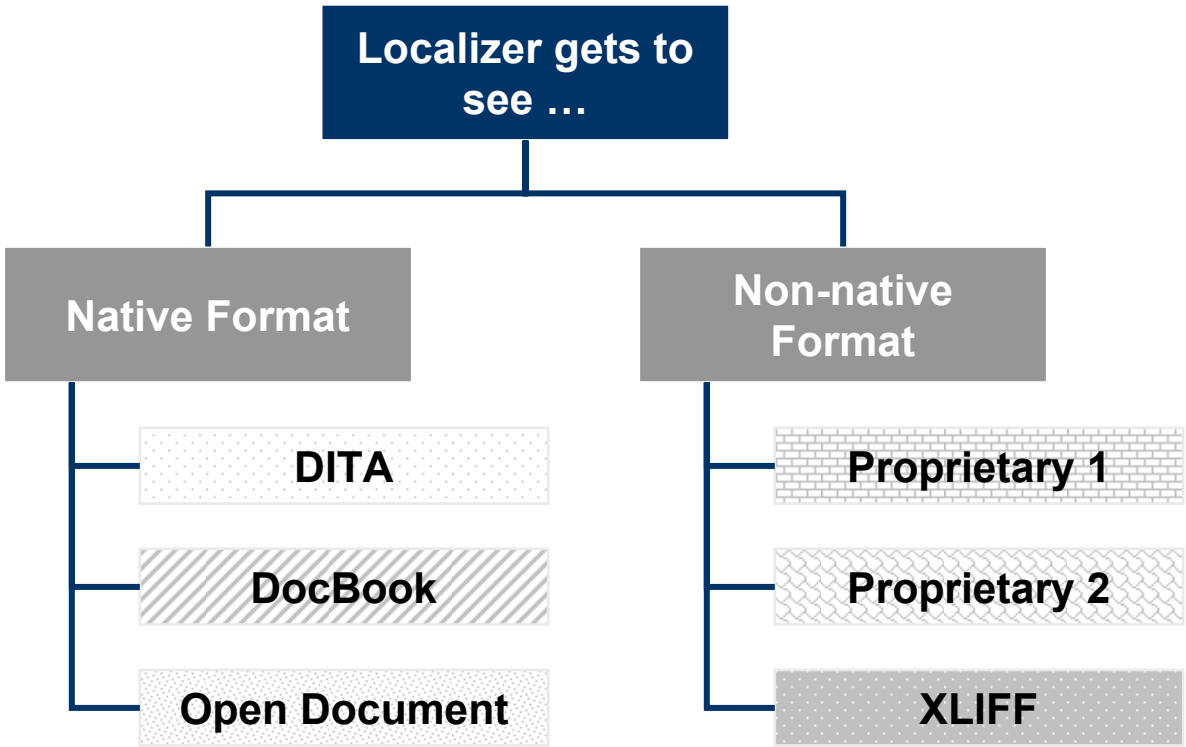
ITS and OASIS Source Formats

ITS and Localization

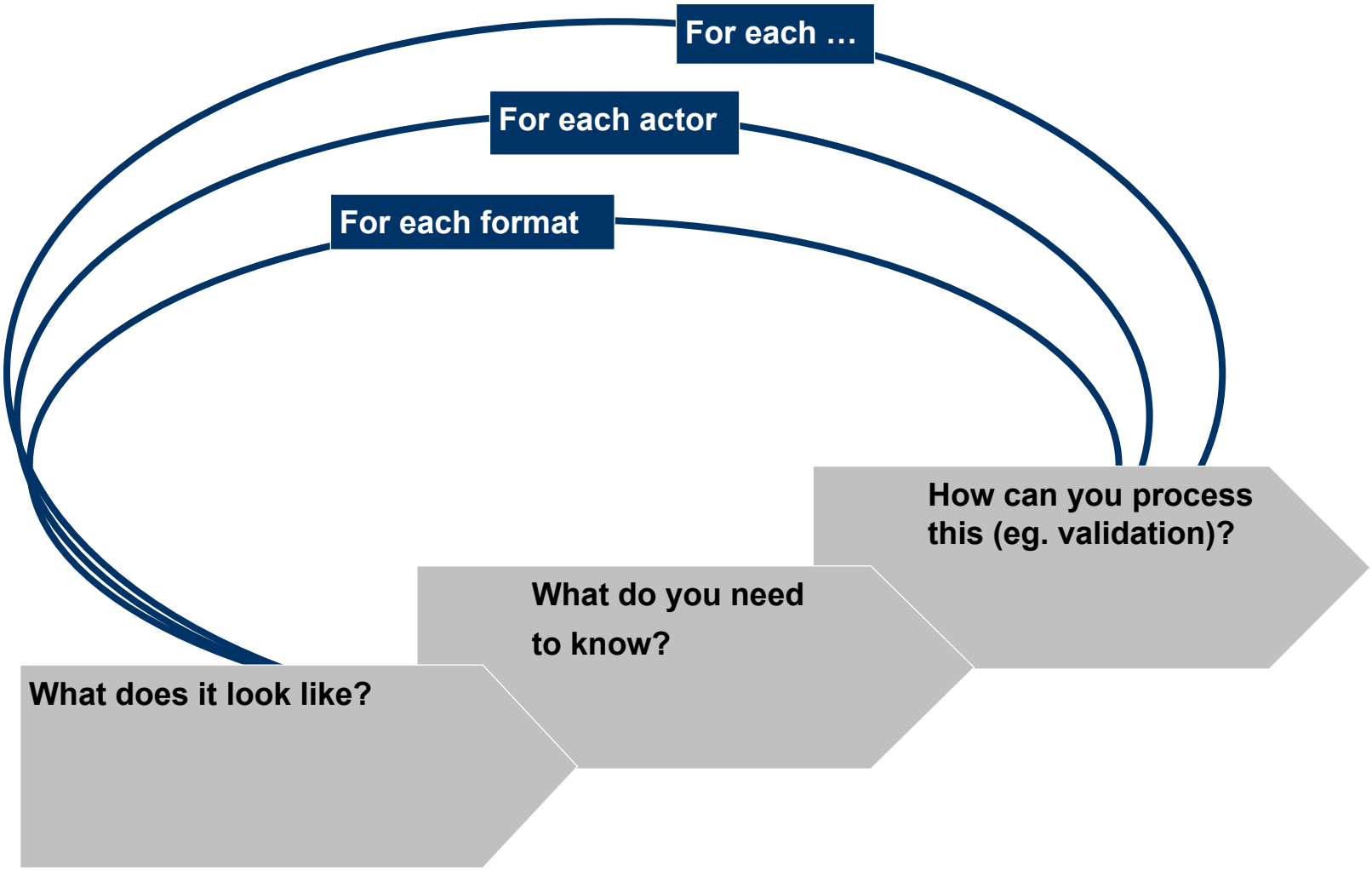
Globalization/Translation Processes – Actors



Globalization/Translation Processes – Formats



Globalization/Translation Processes – The Nightmare



ITS, Localization and Native Format – DITA without ITS

- Language of the content?
- Special content parts?
- Special script markup?
- Markers and linguistic integrity?
- Non-translatable content?
- Annotations for reader?

```
<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 is the same as
       an <b>Appaloosa</b>.</fn> have spotted
     coats. The <term>Nez Perce</term> Indians
      have been key in breeding this type of
      horse.</p>
   </li>
  </ol>
 </conbody>
</concept>
```

ITS, Localization and Native Format – DITA with ITS

Language of the content?

Special content parts?

Special script markup?

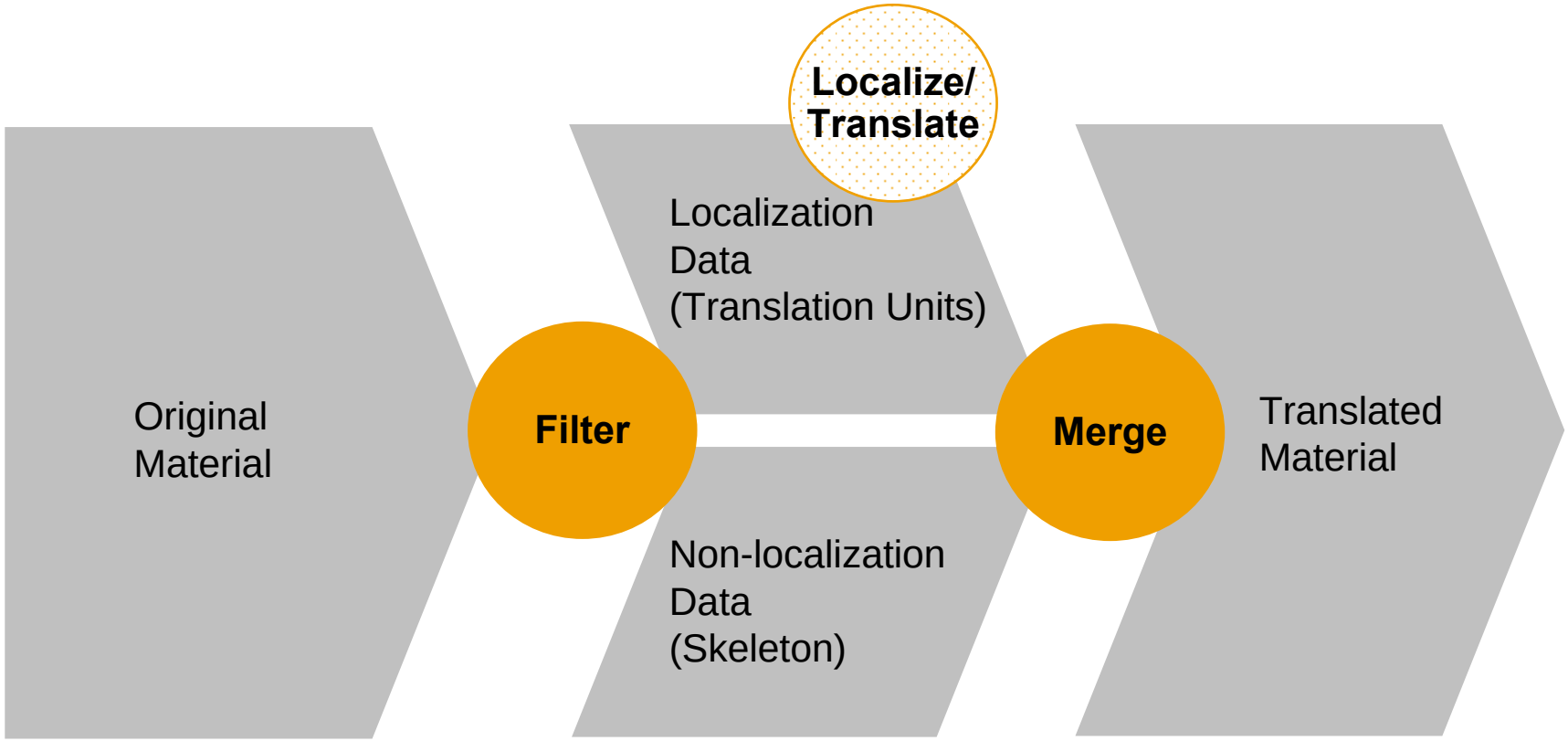
Markers and linguistic integrity?

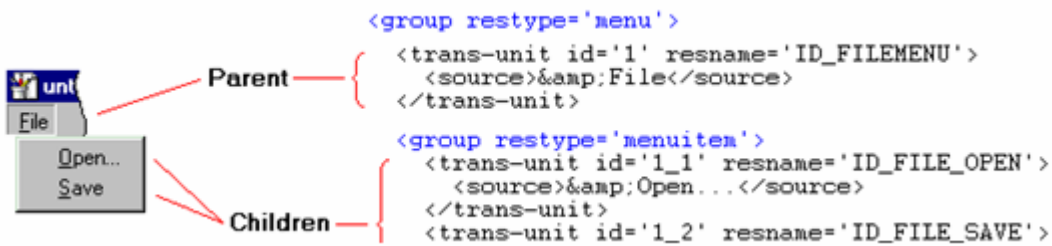
Non-translatable content?

Annotations for reader?

```
<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



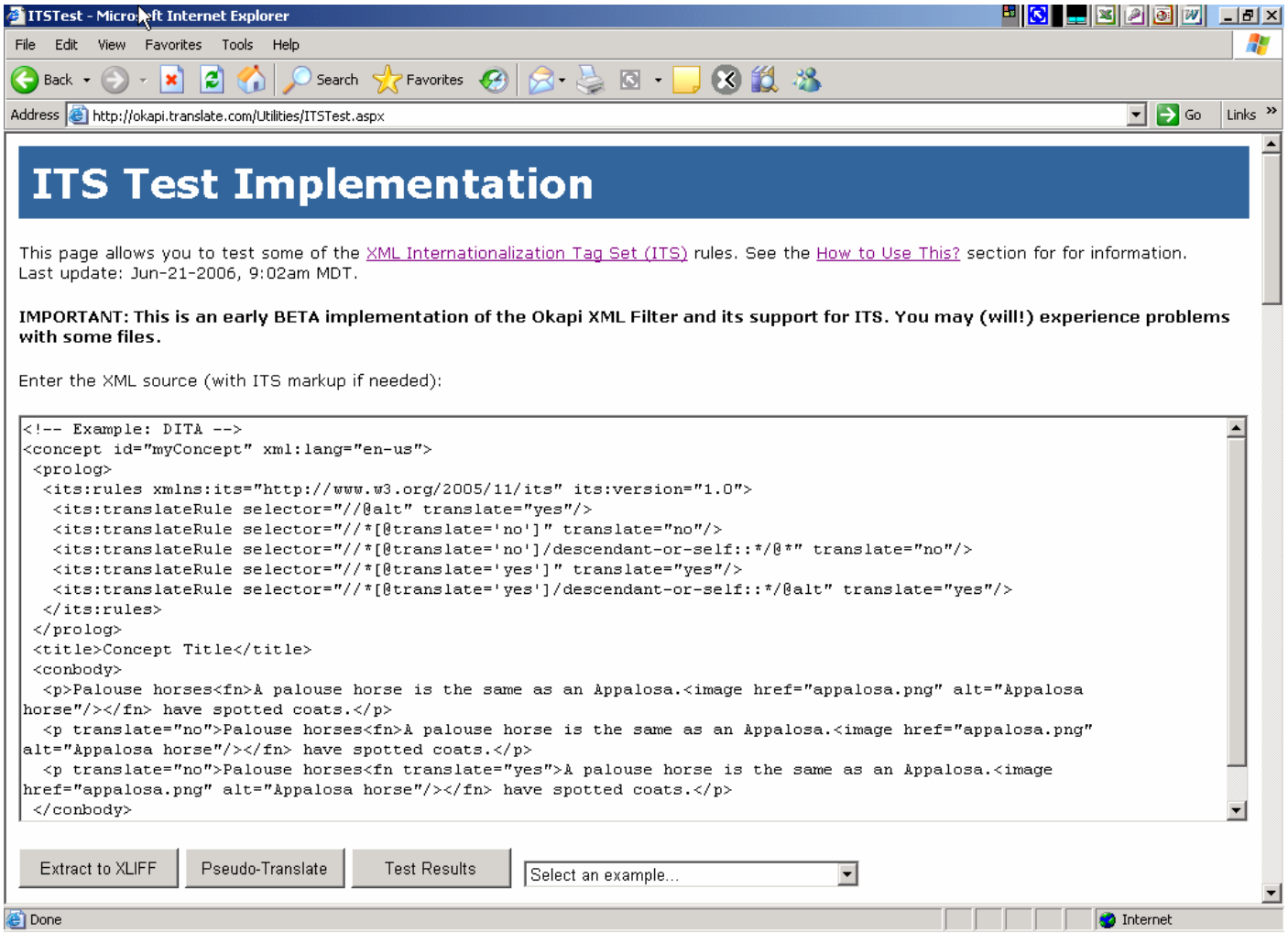


ITS and Interchange Format – Easy Filtering (1/2)

Yves Savourel, ENLASO Corporation

http://okapi.translate.com/Utilities/ITSTest.aspx

Recording



ITS and Interchange Format – Easy Filtering (2/2)

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

Summary (2/3)

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

Summary (3/3)

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