



TRINITY COLLEGE DUBLIN

COLÁISTE NA TRÍONÓIDE

THE  
UNIVERSITY  
OF DUBLIN

# International Tag Set (ITS) 2.0

MultilingualWeb-LT - W3C Working Group

**Dominic Jones**

**MLODE 2012, Leipzig, Germany.**

Chairs: Felix Sasaki, David Filip, David Lewis

# MultilingualWeb-LT

- W3C Working Group under I18n Activity
- **Aims:** define meta-data for web content that facilitates its interaction with language technologies and localization processes.
- Roughly have 45 participants from 25 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.
  - At present ITS (2.0) would add: Domain, Diasambiguation, Locale Filter, Proveance, Text Analysis Annotation, Localisation Quality Issue, Localisation Quality Precip, Terminology, Preserve Space, ID Value, Language Information, Target Pointer, External Resource.
  - ITS 2.0 inherits all 1.0 data categories.

# 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

# ITS 2.0 basics

## **Aims:**

- Rapid localisation of content.

## **Target Content:**

- MLW-LT targets XML, HTML5 and CMS-based 'deep web' content

## **ITS Rules Basics:**

- Embedded or linked Global rules apply to a whole file.
- Local rules applies to the tag at a content level.
- The last processed rule applies to the content.
- Inheritance applies to certain data categories.
- W3C XLINK is for linking to rules in XML and HREF in HTML5.
- Provide the standard, example files, test implementations.

## **Examples:**

- All examples are taken from the W3C working draft located at:  
<http://www.w3.org/TR/2012/WD-its20-20120829/>

# A word of warning...

These examples are from a combination of a **working draft** and **developing test suite**.

They are presented to provide an **overview** of the standard as is today.

They are in no way **finalised** other than as being part of a working draft!

# SMT Metadata

- **Translate:** Expresses whether content is either translatable (translate = yes) or not translatable (translate = no).

```
<messages
  xmlns:its="http://www.w3.org/2005/11/its"
  its:version="2.0">
  <msg num="123">Click Resume Button on Status Display or
    <panelmsg
      its:translate="no">CONTINUE</panelmsg> Button
    on printer panel</msg>
</messages>
```

Example 26: The Translate data category expressed locally  
Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# SMT Metadata

- **Domain:** Identify the domain of content. Does not propose a list of “domains” rather provides a mechanism to associate any values with a section of content.

```
▼<doc>
  ▼<head>
    ▼<its:rules xmlns:its="http://www.w3.org/2005/11/its" version="2.0">
      <its:domainRule selector="/doc/body" domainPointer="/doc/head/meta[@name='DC.subject']/@content"/>
    </its:rules>
    <meta name="DC.subject" content="automotive"/>
  </head>
  ▼<body>
    <p>The new car from XYZ - buy it now!</p>
  </body>
</doc>
```

Domain1Xml.xml

Taken from <http://phaedrus.scss.tcd.ie/its2.0/its-testsuite.html>

# Localizer Signals

- **LocNote:** Provides Translators with information about how to translate content, expand on meaning, clarify ambiguity, indicate emphasis.

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8"><meta>
    <title>LocNote test: Default</title>
  </head>
  <body>
    <p>This is a <span its-loc-note="Check with terminology engineer" its-loc-note-type="alert">motherboard</span>.</p>
  </body>
</html>
```

Example 33: The Localization Note data category expressed locally in HTML5

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **Locale Filter:** Specifies what is applicable on a locale basis. For example: Legal notices for certain regions only. Drop editorial notes from localized content.

```
<book
  xmlns:its="http://www.w3.org/2005/11/its" >
  <info>
    <legalnotice
      its:localeFilterList="en-CA, fr-CA">
        <para>This legal notice is only for Canadian locales.</para>
      </legalnotice>
    </info>
  </book>
```

Example 57: The Locale Filter data category expressed locally  
Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **Directionality:** Allows for specification of base writing direction of text. (ltr, rtl, lro, rlo).

```
<text
  xmlns:its="http://www.w3.org/2005/11/its"  xml:lang="en"
  its:version="2.0">
  <body>
    <par>In Arabic, the title <quote xml:lang="ar"
      its:dir="rtl">W3C نشاط التريل</quote>
      means <quote>Internationalization Activity, W3C</quote>.</par>
  </body>
</text>
```

Example 41: The Directionality data category expressed locally  
Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **External Resource:** Indicates a node represents or references external data outside document such as images, audio or video.

```
<doc
  xmlns:db="http://docbook.org/ns/docbook"
  xmlns:its="http://www.w3.org/2005/11/its" >
<its:rules version="2.0">
  <its:externalResourceRefRule selector="//dbk:imagedata | //dbk:audiodata | //dbk:videodata" externalResourceRefPointer="@fileref"/>
</its:rules>
<db:mediaobject>
  <db:videoobject>
    <db:videodata fileref="movie.avi"/>
  </videoobject>
  <db:imageobject>
    <db:imagedata fileref="movie-frame.gif"/>
  </imageobject>
  <db:textobject>
    <db:para>This video illustrates the proper way to assemble an inverting time distortion
      device. </para>
    <db:warning>
      <db:para> It is imperative that the primary and secondary temporal couplings not be
        mounted in the wrong order. Temporal catastrophe is the likely result. The
        future you destroy may be your own. </para>
    </warning>
  </textobject>
</mediaobject>
</doc>
```

Example 58: The externalResourceRefRule element

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **Ruby:** “Ruby text is used to provide a short annotation of the associated base text. It is most often used to provide a reading (pronunciation) guide.”

```
▼<ex:p>
  Another ruby example:
  ▼<ex:ruby>
    <ex:rb>WWW</ex:rb>
    <ex:rp>(</ex:rp>
    <ex:rt>World Wide Web</ex:rt>
    <ex:rp>)</ex:rp>
  </ex:ruby>
</ex:p>

▼<ex:p>
  A ruby example:
  ▼<ex:ruby>
    ▼<ex:rbc>
      <ex:rb>10</ex:rb>
      <ex:rb>31</ex:rb>
      <ex:rb>2002</ex:rb>
    </ex:rbc>
    ▼<ex:rtc>
      <ex:rt>Month</ex:rt>
      <ex:rt>Day</ex:rt>
      <ex:rt>Year</ex:rt>
    </ex:rtc>
    ▼<ex:rtc>
      <ex:rt rbspan="3">Expiration Date</ex:rt>
    </ex:rtc>
  </ex:ruby>
</ex:p>
```

EX-ruby-pointer-1.xml

Taken from <http://phaedrus.scss.tcd.ie/its2.0/its-testsuite.html>

# Localizer Signals

- **Target Pointer:** Identifies different sub-language pieces of content in the same file.

```
▼<file xmlns:its="http://www.w3.org/2005/11/its">
  ▼<its:rules version="2.0">
    <its:translateRule selector="/file" translate="no"/>
    <its:translateRule selector="//source" translate="yes"/>
    <its:targetPointerRule selector="//source" targetPointer="../target"/>
  </its:rules>
  ▼<entry id="one">
    <source>Remember last folder</source>
    <target>lastFolder</target>
  </entry>
  ▼<entry id="two">
    <source>Custom file filter:</source>
    <target>testFile.txt</target>
  </entry>
</file>
```

targetpointer1xml.xml

Taken from <http://phaedrus.scss.tcd.ie/its2.0/its-testsuite.html>

# Localizer Signals

- **Language Information:** Add BCP47 codes to text identifying language. Expresses the language of a given piece of content.

```
▼<text language="en">
  ▼<head>
    ▼<its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0">
      <its:langRule selector="//*" langPointer="@language"/>
    </its:rules>
  </head>
  ▼<content>
    ▼<par>
      The motto of Québec is:
      <q language="fr-CA">Je me souviens</q>
    </par>
    ▼<par>
      The one of Canada:
      <q language="la">A mari usque ad mare</q>
    </par>
  </content>
</text>
▼<!--
  timestamp $Id: Language1.xml,v 1.4 2007/02/02 16:55:05 srahtz3 Exp $
-->
```

Language1.xml

Taken from <http://phaedrus.scss.tcd.ie/its2.0/its-testsuite.html>

# Localizer Signals

- **ID Value:** “The Id Value data category indicates a value that can be used as unique identifier for a given part of the content.”

```
<resources
  xmlns:its="http://www.w3.org/2005/11/its" >
  <its:rules version="2.0">
    <its:translateRule translate="no" selector="/resources"/>
    <its:translateRule translate="yes" selector="//text"/>
    <its:idValueRule selector="//text" idValue="../@name"/>
  </its:rules>
  <entry name="btn.OK">
    <text>OK</text>
    <pos>1, 1</pos>
    <trig>sendOK</trig>
  </entry>
  <entry name="btn.CANCEL">
    <text>Cancel</text>
    <pos>2, 1</pos>
    <trig>cancelAll</trig>
  </entry>
</resources>
```

Example 62: Pointing to an ID value with the idValueRule element

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **Elements within Text:** Breaks a single text down into a number of elements. Allows for the re-use of certain elements and text segmentation.

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8"><meta>
    <title>Within text test: Default</title>
  </head>
  <body>
    <p>Text with <span its-within-text="yes">bold</span>.</p>
  </body>
</html>
```

Example 48: The Elements Within Text data category expressed locally in HTML5

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localizer Signals

- **Preserve Space:** Indicate whether you want preserve space or not. How should whitespace be handled? Only available in XML.

```
<book
  xmlns:its="http://www.w3.org/2005/11/its" >
  <info>
    <its:rules
      its:version="2.0">
        <its:preserveSpaceRule selector="//verse" space="preserve"/>
      </its:rules>
    </info>
    <verse>
      'Twas brillig, and the slithy toves
      Did gyre and gimble in the wabe;
      All mimsy were the borogoves,
      And the mome raths outgrabe.
    </verse>
  </book>
```

Example 65: The Preserve Space data category expressed globally

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Terminology

- Indicates terminology associated with the text. For example meaning of the text its origin and definition.

```
<text
  xmlns:its="http://www.w3.org/2005/11/its" >
  <its:rules version="2.0">
    <its:termRule selector="//term" term="yes" termInfoPointer="id(@def)"/>
  </its:rules>
  <p>We may define <term def="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>
```

Example 34: Usage of the termInfoPointer attribute

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Disambiguation

- Provides a way of informing an MT or cms systems with more information about a translation.

```
<text
  xmlns:its="http://www.w3.org/2005/11/its" >
  <its:rules version="2.0">
    <its:disambiguationRule
      selector="/text/body/p/*[@id='dublin']"
      entityTypeSourceRef="http://nerd.eurecom.fr/ontology"
      entityTypeRef="http://nerd.eurecom.fr/ontology#Place"
      disambigType="entity"
      disambigSourceRef="http://dbpedia.org/"
      disambigIdentRef="http://dbpedia.org/resource/Dublin"/>
    <its:disambiguationRule
      selector="/text/body/p/*[@id='capital']"
      disambigType="lexicalConcept"
      disambigSourceRef="http://www.w3.org/2006/03/wn/wn20/rdf/wordnet-synset.rdf"
      disambigIdentRef="http://www.w3.org/2006/03/wn/wn20/instances/worsense-capital-noun-3"/>
  </its:rules>
  <body>
    <p>
      <span id="dublin">Dublin</span> is the <span id="capital">capital</span> of Ireland</p>
    </body>
  </text>
```

Example 51: Usage of entityTypeSourceRef, enttiyTypeRef, disambigSourceRef, disambigIdentRef for both entity and word sense disambiguation. Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localisation Quality Issue

- The Localization Quality Issue data category is used to express information related to localization quality assessment tasks. This data category can be used in a number of ways, including the following example scenarios:
  - An automatic quality checking tool flags a number of potential quality issues in an XML or HTML file and marks them up using ITS 2.0 markup
  - A quality assessment process identifies a number of issues and adds the ITS markup to a rendered HTML preview of an XML file along with CSS styling that highlights these issues.
  - A human reviewer working with a web-based tool adds quality markup, including comments and suggestions, to a localized text as part of the review process.

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

# Localisation Quality Précis

- An overall measurement of the localization quality of a document.

```
<doc
  xmlns:its="http://www.w3.org/2005/11/its"
  its:version="2.0"
  its:locQualityPrecisScore="100"
  its:locQualityPrecisThreshold="95"
  its:locQualityPrecisProfileRef="http://example.org/qaModel/v13">
<title>The Adventures of Tom Sawyer</title>
<para>He got home pretty late that night, and when he climbed cautiously in at
  the window, he uncovered an ambushade, in the person of his aunt; and when she
  saw the state his clothes were in her resolution to turn his Saturday holiday
  into captivity at hard labor became adamantine in its firmness.</para>
</doc>
```

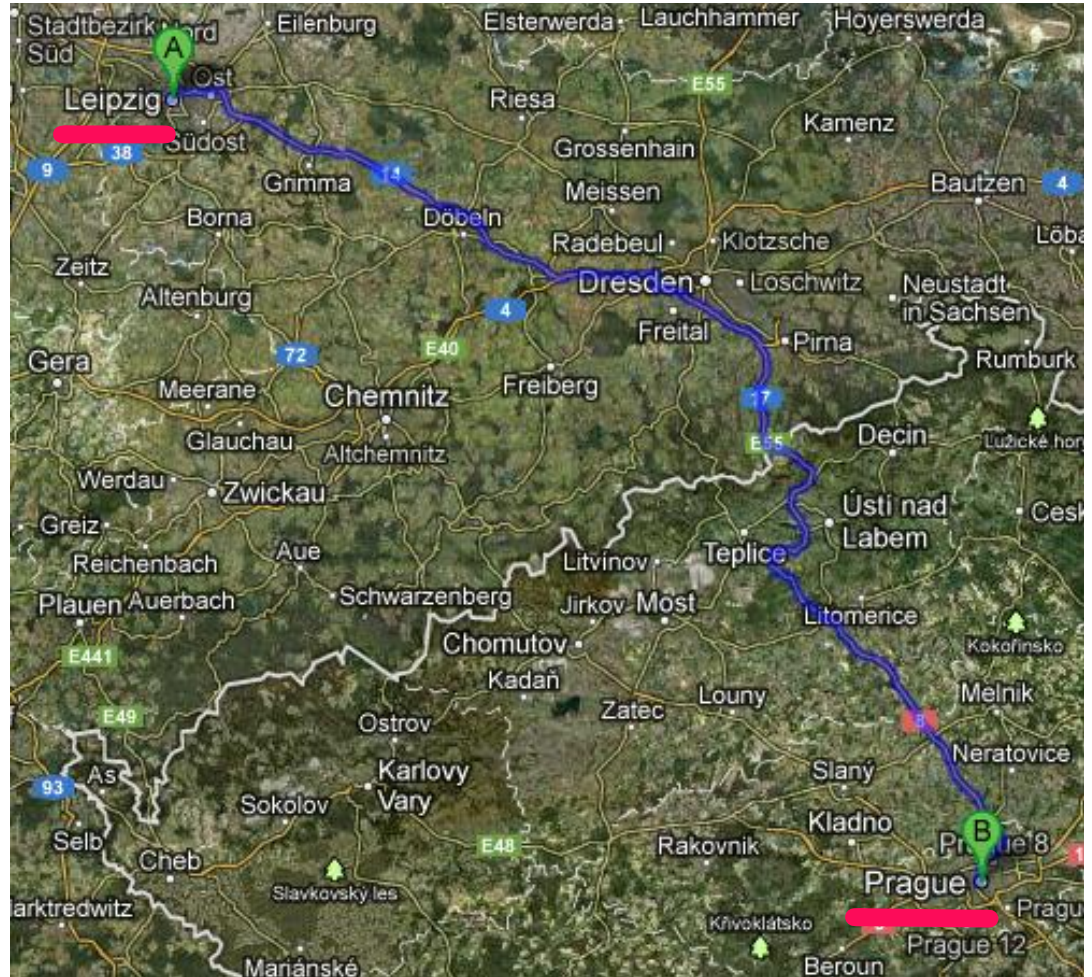
Example 80: The Localization Quality Précis data category expressed locally in XML

Taken from <http://www.w3.org/TR/2012/WD-its20-20120829/>

- Score: Higher values indicating a better score.
- Threshold: Lowest score value that constitutes a passing score

# Currently Going on in Prague...

- Provenance.
- Text Analysis.
- Annotation.
- Test Suite
- Demo Systems
- Implementations
- Czech Beer...



# Its Your Multilingual Web

- Get involved in MultilingualWeb-LT WG
  - Participate as W3C members
  - Follow via public list and WG site
    - <http://www.w3.org/International/multilingualweb/lt/>
  - Offer Implementations
- We're still looking for your input. However:
  - NOVEMBER 2012 = FEATURE FREEZE...**
  - So get in touch... Sooner rather than later!**