State of MLW-LT Requirements June 2012

Felix Sasaki

Dave Lewis

Birds eye view – goal to standardize ITS "data categories"

- Definitions in prose: "Translate data category is used to identify translatable or non translatable pieces of content. It has the values 'yes' or 'no'."

- Implementation in various ways
  - "Locally" XML: ITS 1.0 "its:translate" attribute
  - "Locally" HTML5: "translate" attribute
  - "Global rules" XML / HTML5: rules using XPath expressions – e.g. "//code" (all code elements)
  - Other types of content: within CMS (potentially "just HTML5"), RDFa, microdata, others to be confirmed

- How local "Translate" it already used – example from online MT systems

http://www.w3.org/International/tests/html-css/translate/results-online

Background: ITS 1.0 vs. ITS 2.0

- Focus on XML
- Our Requirements are for ITS 2.0
  - Other types of content, e.g. HTML5
  - Other serializations of data categories, e.g. RDFa, microdata, ...
  - More data categories for localization, language technologies

ITS 1.0 data categories

- Translate
  - Identify whether a piece of content should be translated or not
- Localization Note
  - Human readable notes for localizers
- Terminology
  - Identify terms and add additional human readable information
- Directionality
  - Specify the direction of text
- Ruby
  - Add additional information e.g. for pronunciation
- Language Information
  - Specify language of the document
- Elements within Text
  - Identify nesting (e.g. "span") or block level (e.g. "p") for elements

ITS 2.0 Data categories with rough consensus

- Disambiguation
  - Data for MT systems in disambiguation processes, e.g. pointer to concept reference, thesauri, controlled vocabularies etc.
- ElementsWithinText
- Language information
- Locale filter
  - indication of source content elements as only being suitable for inclusion for localisation to specific locales only
- Localization note
- TextAnalysisAnnotation
  - Information about tools and confidence of annotation
- Translate

Current state of Requirements

- General approach for defining data categories decided, see "representation formats" session
- A few data categories have rough consensus to be added to ITS 2.0, see last slide and http://tinyurl.com/its2-datacats-consensus
- Many data categories have implementation interest, see http://tinyurl.com/its-implementation-proposals
- Now it's time to make decisions!

Next steps

- Decide about what data categories to do and what not by the end of July

- Stable definitions of data categories by end of October

- Stable ITS 2.0 draft by end of November

… oh boy, so much to do – we need your help!

Next steps for underlying EU project

- By mid July: prototype "Translate" + "your most favourite data category" – which one to be decided <u>in this meeting</u>!
  - At least in HTML5 and general XML; at least locally; better: also globally
  - With your customers / data / testing / …
- For others FIY - prototype "where"? See deliverables from WP3-5 at http://www.w3.org/International/multilingualweb/lt/wiki/Deliverables
- By end of July: fix set of data categories in the working group
- By end of September: gather more implementation experience and stabilize the data category definition

This meeting

- Get your ideas about what to do and what not to do

- Gather interest in data categories and implementations

- Get ideas about making money ☺ - what customer might be interested in which data category?

- Don't worry about details of our discussion so far – the session chairs will guide you through these

Q/A