Chapter 4

Metadata for the multilingual web

Felix Sasaki

DFKI and W3C Fellow

We describe the Internationalization Tag Set (ITS) 2.0, an upcoming standard to foster the development of the multilingual Web. ITS 2.0 provides metadata to integrate workflows for content production, localization and language technology. The technical goal is to achieve better results in content creation and other language-related processes; the goal in terms of community building is to raise awareness of needs in multilingual workflows. This aim is also supported by providing re-usable software components for various use cases.[1]

1 Introduction

Content in languages other than English is growing on the Web. But so far a lot of content resides in "language silos". A study by Ford & Batson (2011) reveals that Web pages rarely have links to other languages even of neighbouring countries. Also, the links to English web pages are rather few. This demonstrates that English has not developed into the "lingua franca" of the Web. This has a huge economic impact. A *Flash Eurobarometer* (2011) study indicates for example that 51% of European retailers sell via the Internet, but only 21% support cross-border transactions.

The situation of language silos is also given on the Semantic Web. Ell et al. (2011) have analysed human-readable labels in the Semantic Web. Less than 5% of Uniform Resource Identifiers (URIs) have a language tag, and less than 1% contain labels in several languages. One might argue that in the Semantic Web human readable labels are not needed. But to query the Semantic Web across

[1] The work described in this paper was funded by the European Commission (project name MultilingualWeb-LT) through the Seventh Framework Programme (FP7) Grant Agreement No. 287815.

Felix Sasaki. Metadata for the multilingual web. In Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt (eds.), *Language technologies for a multilingual Europe: TC3 III*, 43–53. Berlin: Language Science Press. DOI:10.5281/zenodo.1291928

languages, query authors need to work with labels or inter-language links leading to resources in their own languages; otherwise non-Japanese speakers, for instance, cannot make use of URIs like http://ja.dbpedia.org/page/講談社 to formulate adequate queries across languages in the Semantic Web.

Translation and creation of cross-language links between (Semantic) Web resources can improve the situation. The challenge here is scalability and cost. Language technology like cross-lingual search and machine translation has gained widespread adoption (e.g. as part of search-engine interfaces). But the translation quality often is rather poor, especially if "distant" languages like German and Japanese are processed, or languages with smaller speaker communities are in scope. As Kornai (2012) discusses, such languages rarely have a lobby on the Web: they lack basic language resources for creating multilingual applications and might even face a "digital extinction".

This paper explores how standardization can help to address challenges faced by the multilingual Web. The upcoming standard "Internationalization Tag Set (ITS) 2.0"[2] fills a gap that hinders better quality in translation on the Web: the availability of metadata to influence multilingual content authoring, translation and localization workflows, using humans and/or language technology.

2 Background

2.1 The MultilingualWeb community

The standardization of ITS 2.0 has emerged from the MultilingualWeb project[3]. Funded by the European commission and lead by the W3C (World Wide Web Consortium), the project started in 2010 with two aims. First, MultilingualWeb brings together stakeholders who are interested in the multilingual Web: language technology researchers, localization service providers, Web technology developers and standardization experts, users from various communities and policy makers who support various regions and their linguistic diversity.

Second, MultilingualWeb has the aim of detecting gaps that hinder the adoption of the multilingual Web. The focus here is gaps related to standardization. Since MultilingualWeb is lead by the W3C, which is the main provider of Web technology standardization building blocks, MultilingualWeb is in a good position to discuss standardization related gaps and to help closing these.

[2] The latest draft of ITS 2.0 is available at http://www.w3.org/TR/its20/ The predecessor ITS 1.0 is available at http://www.w3.org/TR/its/

[3] See http://multilingualweb.eu/ for further information.

MultilingualWeb is running workshops as the main instrument to achieve its goals. Since the start of the first underlying EU project, the EU thematic network MultilingualWeb, four workshops have taken place. Due to the success of the workshops, the MultilingualWeb brand was continued: the successor project called MultilingualWeb-LT (MLW-LT)[4] is supporting the standardization of ITS 2.0 within W3C and the continuation of the MultilingualWeb workshop series and its community. The creation of the MLW-LT EU project and the related W3C group working on ITS 2.0 was a direct result of community building at MultilingualWeb workshops.

2.2 Metadata for the MultilingualWeb: A simple example

At the MultilingualWeb workshops, the topic of metadata for supporting multilingual content creation and related processes came up frequently. Some metadata items like language or character encoding information have been in use for quite some time and are available in various parts of the Web architecture, e.g. HTML Web content or HTTP server settings. One concrete metadata item has been lacking for a long time: a means to identify pieces of content as non-translatable.

Such translation metadata is useful both for language technology, i.e. machine translation systems, and human translators. A standardized means to convey the metadata can ease the creation of high quality localization workflows. The metadata is created by content producers in one language, taken up by localization service providers, and brought to various (human) translators. Here the metadata helps to create a better translation result.

The predecessor of ITS 2.0, that is ITS 1.0, provides a "Translate" metadata item. Metadata items in ITS 1.0 and ITS 2.0 are so-called "data categories". Discussion about adding a "translate" attribute implementing the "Translate" metadata category in HTML5 started in 2008; the attribute eventually was added to the HTML5 draft in 2012. The MultilingualWeb community helped significantly to raise awareness about the topic, see e.g. the presentation of Ishida & Kosek (2011).

2.3 From "Translate" to enhanced metadata

Soon after adding the attribute to HTML5, two online machine translation services provided support: Bing Translator and Google Translate. [5] This demonstrated the usefulness of metadata for multilingual Web content processing.

[4] See http://www.w3.org/International/multilingualweb/lt/ for further information.

[5] Test results and example files demonstrating the functionality are available at http://www.w3.org/International/tests/html-css/translate/results-online

However, the “Translate” data category is only the tip of the iceberg: already ITS 1.0 provides further data categories like “Terminology” markers for terms, “Elements within Text” indicators of nested text flows (e.g. embedded footnotes) and others.

The scope of ITS 1.0 is XML content; for ITS 2.0, the aim is to provide the data categories also for HTML5 or other flavours of HTML. In addition, ITS 2.0 provides further data categories that support workflows between Web content authoring environments, language technology applications and localization tools.

3 Introduction to ITS 2.0

3.1 Basic principles

Both ITS 1.0 and ITS 2.0 share the same basic principles. Metadata items, that is the “data categories”, are defined independently of their usage or “implementation”. An example is the “Translate” data category. Its purpose is to convey two kinds of information: a piece of content is translatable or not. The implementation of “Translate” can happen via a “translate” attribute as in HTML5. Adding ITS markup directly into a document is called the ITS “local approach”.

In many workflows, data categories are not set by content creators locally for each piece of information. The metadata is rather introduced by information architects working on a document format or project template basis. For this scenario, ITS provides an XML approach of “global rules”. The following ITS file contains a rule demonstrating this functionality for the “Translate” data category.

```
<its:rules ...>
  <its:translateRule translate="no" selector="//code"/>
</its:rules>
```

The “its:rules” element serves as a wrapper. The “its:translateRule” element contains a “selector” attribute. Via an XPath expression, all “code” elements are selected. The “translate” attribute set to “no” expresses that these elements should not be translated.

ITS global rules are independent of a given document, that is: what “code” elements are matched depends on the actual content being processed.

In addition to global rules and local markup, ITS provides further data category specific definitions, like inheritance behaviour of ITS information (e.g. inheriting “Translate” information to child elements of selected element nodes) or defaults. For example the default for “Translate” is that elements are translatable and attribute values are not translatable.

3.2 Types of content: from XML to HTML

As described above, ITS 1.0 was defined with a focus on XML content. This raises the question how XML specific technologies like XPath can be used to process other types of Web content. A few years ago the focus of web technology development was on XHTML, the XML version of HTML. Today HTML5 needs to be taken into account. It provides an XML form too, but also a widely used, non-XML serialization.

The ITS 2.0 approach to accommodate this development has four aspects. First, data categories that are available natively in HTML are mapped to ITS 2.0 definitions, so that an ITS 2.0 processor can take the HTML markup into account. This approach is taken e.g. for the "Translate" data category and the "Language Information"data category, which conveys language information in the same way as the HTML "lang" or XHTML "xml:lang" attributes.

Second, ITS 2.0 provides counterparts of ITS local markup in a manner that easily can be integrated into Web content. The below example shows local ITS markup for "Terminology" information in an arbitrary XML format, using a "term" attribute in the ITS namespace.

```
<p ...>
 And he said: you need a new
 <quote its:term ="yes">motherboard</quote>
</p>
```

The HTML counterpart replaces the XML namespace mechanism with a hardwired prefix `its-*`.

```
<p ...>
 And he said: you need a new
 <quote its-term="yes">motherboard</quote>
</p>
```

The HTML validation service validator.nu,[6] which is the basis for the HTML5 part of the W3C markup validator, already provides a preset (HTML5 + SVG1.1 + MathML3.0 + ITS2.0) for validating this kind local ITS 2.0 in HTML5 markup.

Third, to be able to re-use global rules with various serialization flavours of HTML5, ITS 2.0 foresees a processing chain that takes the serializations as input and creates one common DOM (document object model) in memory representation. This representation can be processed with XPath. The output then can be serialized into different forms. The aforementioned validator.nu service provides an HTML5 parser to realize both the DOM generation and the output serializations.

[6] See http://validator.nu/ for more information.

Finally, in ITS 2.0, the selection mechanism of global rules, that is XPath, can be replaced by CSS selectors. Various libraries to convert CSS selectors into XPath expressions exist; in this manner, content authors and content managment system (CMS) template editors can use the selectors technology of their preference and convert the CSS selectors into XPath before actual processing. This approach helps to make ITS data categories accessible for a wide range of users.

3.3 A birds eye view on ITS data categories

ITS 1.0 provides data categories with a focus on two areas. The first is translation and localization processes. “Translate” or “Term” are examples of relevant data categories. The second area is called “internationalization”. In ITS 1.0, internationalization related data categories encompass metadata needed for content authoring in specific cultural or language regions. The main data categories here are: “Ruby”, used to add among others pronunciation information to texts e.g. in the Japanese script; and “Directionality”, used to specify the base writing direction for e.g. the Arabic or Hebrew script.

In ITS 2.0, localization related data categories are being extended and language technology related metadata is provided. An example for new localization related data categories is “Locale Filter”. It identifies content that is relevant (or not relevant) for a given locale. “Allowed characters” defines characters that are permitted to appear in a piece of content, e.g. in certain parts of a user interface.

Language-technology related data categories help to create workflows including e.g. machine translation process. An example here is “Domain”, see the following `its:domainRule` element.

```
<its:rules ...>
 <its:domainRule selector=”/h:html/h:body” domain
  Pointer=”/h:html/h:head/h:meta[@name=’keywords’]/ @content” />
</its:rules>
```

The “selector” attribute selects the body of the HTML content via an XPath expression, in the same manner as the selector described above for the “translateRule” element. The “domainPointer” attribute selects keywords available in the HTML content: a certain “meta” element. Such domain information then can be used e.g. by machine translation systems to choose the appropriate subsystem being trained for certain text domains.

Another language-technology related data category is “MT Confidence”. A machine translation system can use it to express confidence information about the translation. For other data categories like “Terminology”, which may be created via automatic annotation processes, such confidence information is provided as well.

4 Metadata versus, for or in linguistic annotation?

Annotating textual content as a resource for language related processing is not new. Linguistic corpora including annotations have been developed for decades. Efforts in a forum like ISO TC 37 / SC4 have led to standards for linguistic annotation. ITS both 1.0 and 2.0 are different with respect to their main focus. They do not focus on adding information about linguistic categories on various levels (e.g. morphology, syntax, semantics) to textual content, but non-linguistic, mostly process related metadata (e.g. start time, end time, CPU seconds used etc.).

However, some data categories for ITS 2.0 have a close relation to linguistic annotations. An example is the aforementioned "Terminology". A data category that has been added to ITS 2.0 is called "Text Analysis". It uses the prefix "its-ta" in HTML. The aim is to represent the output of an automatic annotation process. In the below example it is assumed that the string "Dublin" has been annotated as a result of such a process.

```
<span
  its-ta-confidence=”0.7”
  its-ta-class-ref=”http://nerd.eurecom.fr/ontology#Place”
  its-ta-ident-ref=”http://dbpedia.org/resource/Frankfurt_(Oder)”>Frankfurt
</span>
```

"ta-confidence" provides tool-generated confidence information, similar to "MT Confidence" or confidence information for "Terminology". "ta-class-ref" contains a reference to the class of unit being annotated, here making use of the NERD ontology, see Rizzo et al. (2012). "ta-ident-ref" is a unique identifier of the unit, here taken from the DBpedia structured information source, see Kobilarov et al. (2007).

Making this kind of metadata available beyond the realm of language technology has great promises. Localization workflows can convey information to translators and speed up translation. In the above example, the "its-ta-ident-ref" attribute helps to disambiguate the reference of *Frankfurt* in the given text.

Before providing real value, however, challenges have to be addressed. Some tools may assign different ta-ident-ref attributes to the same unit. This leads to a need for annotating the same content with competing pieces information. Many approaches to realizing this requirement exist[7] – but should ITS 2.0 try to adopt these?

Such topics are currently under discussion. The direction seen on the horizon is along the lines of "divide and conquer": ITS 2.0 will keep the focus on simple inline annotations, providing mostly container attributes for the output of text analysis tools. In case of conflicting information or decisions to be taken about how to categorize concurrent annotations, ITS 2.0 is only a starting point for further linguistic processing.

The decision about what formats are to be used here is out of scope for ITS 2.0. Nevertheless, the current ITS 2.0 draft provides an algorithm to convert ITS 2.0 annotated documents into the NIF format, see Rizzo et al. (2012). Using a NIF wrapper, more complex linguistic processing can take place, and the output can be integrated into ITS 2.0 "ta-*" representations again.

5 Use cases and reference implementations

ITS 2.0 by no means tries to solve all issues of metadata for the multilingual Web. As the previous section has shown, areas like linguistic annotation are rather left to other technology areas and standardization efforts. ITS 2.0 focuses on certain use cases. These also have driven the definition of the standard itself. Below is a short summary of major use cases. Additional information is provided by Lieske (2013).

5.1 Simple machine translation

In this use case, XML or HTML5 documents are translated using a machine translation service. The textual content is extracted based on ITS 2.0 data categories. The extracted content is then sent to the machine translation service. The translated content is finally merged back into the original format.

For this use case, "Translate" and "Locale Filter" are useful data categories. "Elements within Text" helps to drive the extraction process as well, e.g. for separating footnotes from the overall text flow. Another data category is "Preserve space": it helps to assure proper handling of whitespace in the translated text. Depending on the capabilities of the machine translation system, "Domain" information can be taken into account as well.

[7] The TEI provides an overview of these approaches, see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html

5.2 Translation package creation

The aim here is to convert input text into a translation package format like XLIFF. Like in the machine translation use case, ITS 2.0 metadata drives the extraction process. Compared to that use case, additional data categories are taken into account, like “Allowed Characters” or “Terminology”. During the extraction process, the ITS 2.0 metadata is transformed into an XLIFF representation. The actual role of the metadata then depends on the translation tool being used.

5.3 Integration of CMS and TMS systems

Often Web content is created via a CMS. Hence, the integration of a CMS with translation managment systems TMS is a major task for creating localization workflows. In this use case, ITS 2.0 data categories help to streamline the localization workflow.

The same data categories as in the translation package creation are relevant for this use case. The main difference is that no dedicated package format like XLIFF is being used.

5.4 Terminology and text analysis annotation

These use cases encompass the automatic services to create ITS 2.0 annotations described above.

5.5 Reference implementations

The use cases are demonstrated by various reference implementations. These are being developed within the EU project underlying the MLW-LT group. The output mostly will be open source implementations, to foster the widespread adoption of the metadata.

6 Conclusion and future work

This paper described ITS 2.0, an upcoming standard that provides metadata to integrate workflows for content production, localization and language technology. We discussed the MultilingualWeb community whose efforts led to the creation of ITS 2.0. Then we introduced the basic principles of the upcoming standard and technical details.

Various metadata items, so-called “data categories”, are being provided by ITS 2.0. We discussed some of them; the area of text analysis annotation has challenges and promises and may help to apply language-technology based, linguistic annotations within localization tool chains. Finally, we discussed some use cases that demonstrate the application of ITS 2.0 metadata, and reference implementations.

The metadata definitions of ITS 2.0 were finalized during 2013, and reference implementations helped to foster their adoption. The publication of the final ITS 2.0 standard was issued on 29 October 2013.

The work undertaken for ITS 2.0 has focused on basic infrastructure for the multilingual Web. Currently detailed topics of the next decade for research in the area of language technology are being defined. The META-NET Strategic Research Agenda (SRA), described by Rehm in this volume, played a major role in shaping these topics. Among these are areas like multilingual Semantic Web, which has been discussed in the introduction of this paper. One future challenge will be how to use such data from or for the multilingual Semantic Web in localization or language technology applications, while also taking ITS 2.0 metadata into account.

References

Ell, Basil, Denny Vrandečic & Elena Simperl. 2011. Labels in the web of data. In *Proceedings of ISWC 2011*, 162–176.

Flash Eurobarometer. 2011. http://ec.europa.eu/public%5C_opinion/flash/fl%5C_313%5C_en.pdf. User Language Preference Online. Report, May 2011.

Ford, Daniel & Josh Batson. 2011. *Languages of the world (wide web)*. http://googleresearch.blogspot.com/2011/07/languages-of-world-wide-web.html.

Ishida, Richard & Jirka Kosek. 2011. *HTML5 i18n: A report from the front line*. http://www.multilingualweb.eu/en/documents/luxembourg-workshop/luxembourg-workshop-report%5C#ishida.

Kobilarov, Georgi, Piet Hensel, Richard Cyganiak & Christian Bizer. 2007. *DBpedia: A nucleus for a web of open data*. Poster presentation at CSSW07.

Kornai, András. 2012. *Language death in the digital age*. http://www.meta-net.eu/events/meta-forum-2012/report#kornai_presentation. Presentation at META-FORUM 2012, Brussels.

Lieske, Christian. 2013. *Metadata for the multilingual web: usage scenarios and implementations*. http://www.w3.org/TR/mlw-metadata-us-impl/. W3C Working Draft 7 March 2013.

Rizzo, Giuseppe, Troncy Raphaël, Hellmann Sebastian & Bruemmer Martin. 2012. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In Christian Bizer, Tom Heath, Tim Berners-Lee & Hausenblas Michael (eds.), *LDOW, 5th Workshop on Linked Data on the Web*, 1–10. Lyon, France: Linked Data on the Web (LDOW2012).