# Metadata for the Multilingual Web:

# Introducing Internationalization Tag Set (ITS) 2.0

**Felix Sasaki**
DFKI / W3C Fellow
Alt-Moabit 91c
10559 Berlin, Germany
felix.sasaki@dfki.de

## Abstract

Internationalization Tag Set (ITS) 2.0 http://www.w3.org/TR/its20/ is a standard that has been developed within the World Wide Web Consortium (W3C). It provides metadata items ("data categories") that ease the integration of natural language processing into core Web technologies. ITS 2.0 focuses on HTML, XML-based formats in general, and can leverage processing based on the XML Localization Interchange File Format (XLIFF), as well as the Natural Language Processing Interchange Format (NIF).

## 1 Multilingual content production

ITS 2.0 metadata eases multilingual content production by providing metadata for production phases like:

- Internationalization;
- Pre-production, e.g. related to marking terminology;
- Automated content enrichment, e.g. automatic hyperlinking for entities;
- Extraction/filtering of translation-relevant content;
- Segmentation;
- Leveraging, e.g. of existing translation-related assets such as translation memories;
- Machine Translation, e.g. geared towards a specific domain;
- Quality assessment or control of source language or target language content;
- Generation of translation kits, e.g. packages based on XLIFF;
- Post-production; and eventually
- Publishing of multilingual content.

The document "Metadata for the Multilingual Web - Usage Scenarios and Implementations" http://www.w3.org/TR/mlw-metadata-us-impl/ lists 18 different usage scenarios for ITS 2.0. Most of them are composed of several of the aforementioned phases.

## 2 ITS 2.0 scenarios for machine translation

In this poster presentation we will introduce ITS 2.0 data categories that are of high relevance for scenarios involving machine translation. We will also summarize usage scenarios and implementations that demonstrate the role and benefit of the data categories in machine translation workflows.

Finally we will touch upon the most important aspect of ITS 2.0: the communities around language technology, localization and content production in the Web that have been involved in the development of ITS 2.0. The main outcome of the standardization process is the bridges that have been built between these communities, in order to bring language technologies into the Web.

## References

Filip, D., S. McCance, D. Lewis, C. Lieske, A. Lommel, J. Kosek, F. Sasaki and Y. Savourel. *Internationalization Tag Set (ITS) 2.0*. W3C Working Draft 21 May 2013. The latest version of ITS 2.0 is available at http://www.w3.org/TR/its20/.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
*Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 423.*