How the Multilingual Semantic Web can meet the Multilingual Web

A Position Paper

Felix Sasaki
DFKI / W3C Fellow
fsasaki@w3.org

The success of the Web is not based on technology. It is rather based on the availability of tooling to create web content, the fast number of content creators providing content, and finally the users who eagerly "digest" the content and are willing to pay for it, being part of various business models.

Not only the Web in general, but also the Multilingual Web is growing. More and more content is being produced in languages other than English; more and more users want to use their mother tongue on the Web. Unfortunately this growth is not without undesired side effects. "If a language is not on the Web, it doesn't exist" – this phrase[1] expresses the fear of "digital extinction", faced especially by smaller language communities.

To support the Multilingual Web, language technology can play a crucial role: the machine translation of the English Wikipedia articles into Thai is just one example how massive content creation can rely on language technology. The outcome is of course not perfect, and only with human post-editing the result is really useful.

What does all this tell us about the Multilingual Semantic Web (MLSW)? First, like with the Web itself, the availability of standardized technological blocks is a pre-requisite for wide adoption of the MLSW. However, this is not enough. Easy to use tooling to create and to work with RDF based resources is inevitable to lower the barriers for the ordinary content creator. There should be no difference in working with the MLSW compared to editing an HTML web page or setting up a blog.

Second, although the technical infrastructure of the MLSW is given via RDF based building blocks, MLSW resources are rare. Studies[2] reveal that human readable descriptions even in English are hardly available; for other languages or links between languages in the MLSW the situation is even worse.

Third and finally, like for the human readable Web, the application of language technologies can help to create resources for the MLSW, e.g. via the creation of multi-language labels via machine translation. But also like with translation of ordinary Web pages, such approaches need human intervention to assure a certain level of quality.

---

[1] See the presentation from András Kornai at META-FORUM 2012 for details http://www.meta-net.eu/events/meta-forum-2012/report#kornai_presentation

[2] See e.g. the presentation by Jose Emilio Labra Gayo at http://www.multilingualweb.eu/documents/dublin-workshop/dublin-workshop-report#labra

adfa, p. 1, 2011.

The main message of this position statement is that the MLSW has several gaps, which currently hinder the widespread creation and usage of multilingual resources for the Semantic Web. About 2 ½ years ago a similar observation of gaps lead to the creation of a European thematic network, called "The MultilingualWeb"[3]. Via a series of workshops, stakeholders from diverse areas came together and discussed gaps that hinder the development of the Multilingual Web. As one concrete outcome, a EU project was created to develop tooling and standards for a subset of gaps, related to metadata in localization workflows. The W3C MultilingualWeb-LT working group[4] forms the umbrella for this effort. In addition the underlying EU project continues to run the MultilingualWeb workshop series, as a basis for continuous cross-community information exchange and long-term planning.

It seems that for the progress of the MLSW a similar effort is needed. This should not only focus on technology, but on integrating communities. In the remainder of this position statement we will go through the various stakeholder groups identified within the MultilingualWeb workshop series, and will map them to the situation in the MLSW.

Platform Developers provide the technological building blocks that are needed for multilingual content creation and access on the Web. For the Multilingual Web the browser plays a major role. For the MLSW a platform for easy creation of RDF "without seeing the source code" is yet to come. Both the Multilingual Web and the MLSW face challenges in handling of translation workflows. Although more Web content is being translated, the key web technologies HTML and RDF so far have no means to support this process. The forehand mentioned MultilingualWeb-LT working group provides a solution which can be applied to the multilingual Semantic Web as well: upcoming metadata as part of HTML5 based labels in RDF 1.1[5].

The adoption of RDF is hindered by the abstract level of the related standards, lack of outreach, un-harmonized usage of multilingual labeling (see the studies mentioned before), or a lack of testability. A reference implementation of an easy-to-use platform, accompanied by various e.g. educational materials, could boost the adoption of the MLSW. For the Multilingual Web, the W3C has made a long-term effort to raise awareness for multilingual issues via its Internationalization Activity. It is time to work on awareness for the MLSW in this and other fora.

Content Creators more and more need to bring content to different delivery platforms, especially via mobile devices. Since these devices lack computing

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

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

[5] The usage oft the metadata in HTML5 can be seen at http://www.w3.org/TR/its20/#EX-translate-html5 . Since RDF 1.1. encompasses an HTML5 data type, the same approach can be used for translation metadata in RDF labels.

power, many aspects of multilinguality need to be carefully addressed. For the Web in general the creation of applications that work only via the network, e.g. voice analysis and synthesis, has grown. The same holds for the MLSW: device independency can only be achieved if there are stable services which a MLSW "client" can make use of.

The need to create more inter-language links again is valid for both parts of the Web. In the Multilingual Web personalization has become ubiquitous. Search engine providers and other services track user behavior in order to provide the most relevant content in a given situation. The same desire seems to be given for the MLSW: a user e.g. of multilingual RDF resources should not need to have to provide details what parts of the resources (domain specific or general) are relevant; the MLSW "client" should choose the resources based on preferences and tracking of past behavior. Of course such an approach raises privacy issues – and it seems that an initiative like the W3C Tracking Protection working group might then become relevant for the MLSW as well.

The MLSW so far does not address e.g. the requirements of modalities other than text: what role has an image, a video or audio file in the Semantic Web? In the Multilingual Web it is common that such pieces of content are localized to a specific audience – but how about the MLSW? An effort like the English Wikipedia translated into Thai demonstrated the value of combining machine translation with volunteer efforts to create high quality content. For the MLSW, such community approaches are yet to come.

Tooling again seems to be crucial, e.g. to support the easy translation of human readable labels. Explaining the usage of such tooling leads to best practices. For the Web in general, W3C and other organizations recently launched "Web Platform Docs" to provide educational material to a worldwide audience of content creators. Having such material available for the MLSW will be an important step for wider adoption.

Localizers deal with internationalization practice in content creation, the distribution of content to localization companies and the onward distribution to individual translators. Improved efficiency of this process requires technical integration in the resulting workflow.

In this area, the problems of the MLSW are in essence the same as in the Multilingual Web. There is a huge fragmentation of standardization efforts in the localization area. Multiple, sometime overlapping standards are available from different organizations including the W3C, ISO, OASIS, ETSI, or the Unicode consortium. The gap here is often just to understand how the standards interplay.

What does this mean for the MLSW? Truly widespread adoption will mean that Semantic Web resources have to become part of localization workflows and are localized by professional localization companies, by volunteers or a mixture of both. There is no silver bullet to avoid the mistakes being made for

localization of the Multilingual Web. Some advices can be made: not to develop additional standards in this area but rely on existing solutions; integrate localization functionality in a to be developed MLSW platform; and try to match localization workflows, content creators and project needs.

A very promising area in terms of localization tooling seems to be the integration of localization functionality in content creation tools. As mentioned before, the integration of content management and localization is a major task in this area. Bringing MLSW and content creation / localization tooling closer to each other is then just the next logical step.

For **machines**, i.e. applications based on language technology, resources from the MLSW are of (potential or actual) use for cross lingual search, machine translation, multilingual summarization etc. Some language technology applications help to improve the resources of the MLSW, e.g. again machine translation, or data cleansing techniques. The challenges in this area are similar to localization: there are many small solutions, integration has to be done repeatedly, and the re-use of multilingual resources is not straightforward.

Some small integration steps between localization, language technology and the MLSW are being taken. An example is the application of analytics, e.g. named entity annotation, in localization workflows. The dominant format in such workflows is XLIFF (XML Localization Interchange File format). So far there is no standardized way and no tooling available to represent named entities in XLIFF. In the MultilingualWeb-LT working group such tooling is being developed. This will lead to a named entity annotation round tripping workflow from HTML[6] (potentially with an intermediate step via NIF[7]) to XLIFF and back, after translation.

Users normally have no strong voice in the development of multilingual or other technologies. At the MultilingualWeb workshops, it became clear that the worldwide interest in multilingual content is high, but significant organizational and technical challenges need to be tackled for reaching people in less developed economies, especially in linguistically diverse regions such as Asia and Africa. Again, for the creation of content in the MLSW, the same problems apply as well.

A notion that is becoming common in the Multilingual Web is the difference between controlled and uncontrolled environments of content creation and translation. For the MLSW, this seems to be especially crucial for the paradigm of linked open data. Here currently there is practically no difference being made between human language labels created via high quality human translation, or automated results.

---

[6] See an example annotation at http://www.w3.org/TR/its20/#EX-disambiguation-html5-local-1

[7] For details about NIF see http://nlp2rdf.org/nif-1-0

Although the engagement of users is a challenge, it has also promises. A presentation from Facebook at one of the MultilingualWeb workshops revealed that there are 500000 voluntary translators, and that the French instantiation of the site had been translated within 24 hours. A great vision along these lines is a community effort in which fasts amount of content are being created for the MLSW.

Finally, the topic of **policy makers** it is of high importance: many gaps in the Multilingual Web are related to political decisions. Multilingual mandates, participatory democracy or interactive systems for local needs are just a few application scenarios for the MLSW. As a pre-requisite, open multilingual assets are needed, as well as harmonized support across language boundaries. Like with the other areas mentioned in this position statement, such efforts need to be accompanied by education, promotion, coordination, guidelines and business cases related to the MLSW.

Acknowledgements

This position paper was written with input from a paper for the WWW 2012 conference, written by Dave Lewis, David Filip and Felix Sasaki. In addition, a presentation about the outcomes of the MultilingualWeb thematic network given at the TCWorld 2012 conference, given by Christian Lieske and Jan Nelson, provided valuable input. The current MultilingualWeb workshop series receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) Grant Agreement No. 287815.