The Multilingual Web

David Filip
CNGL
University of Limerick
Limerick, Ireland
+353860222158
davidf@ul.ie

Dave Lewis
CNGL
Trinity College Dublin
Dublin 2, Ireland
+35318968428
dave.lewis@scss.tcd.ie

Felix Sasaki
DFKI
Alt-Moabit 91c
10559 Berlin, Germany
+49 30 238951807
felix.sasaki@dfki.de

ABSTRACT
We report on the MultilingualWeb initiative, a collaboration between the W3C Internationalization Activity and the European Commission, realized as a series of EC-funded projects. We review the outcomes of "MultilingualWeb", which conducted 4 workshops analyzing "gaps" within Web standardization that currently hinder multilinguality. Gap analysis led to formation of "MultilingualWeb-LT" – project and W3C Working Group with cross industry representation that will address priority issues via standardization of interoperability metadata.

Categories and Subject Descriptors
H.5.4 Hypertext/Hypermedia

General Terms
Standardization

Keywords
Language Technology, Localization, Internationalization, Web Technologies, Standardization, Metadata, Interoperability

1. MULTILINGUAL WEB: OVERVIEW
MultilingualWeb Initiative (http://www.multilingualweb.eu/) began as an EC-funded thematic network project, exploring standards and best practices supporting the creation, localization and use of multilingual web-based information. It is led by W3C, the major stakeholder for creating the technological building blocks of the web. MultilingualWeb has 22 partners representing research institutes and various industries related to content creation, content localization and associated software vendors (see http://www.multilingualweb.eu/partners). Series of four public workshops took place, from October 2010 to March 2012, in Madrid, Pisa, Limerick and Luxembourg respectively. Their focus was the standards and current best practices to enable a fully multilingual web, and gaps to be filled. They have been of enormous success, in terms of the number of participants, awareness in social media, and the outcome of discussions.

The formation of the W3C Working Group MultilingualWeb-LT, on March 7, 2012 resulted from this success. This WG will develop standardized metadata for Web content to seamlessly interact with language technologies, to enable localization. This will happen with broad and representative consensus under the W3C Internationalization Activity. The WG Charter has been endorsed by 24 W3C members (11 outside the EC-funded group).

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. STAKEHOLDER COMMUNITIES
The initiative recognized the need to connect several communities with different roles in achieving widespread multilingualism on the web. *Internationalization* in content management industry deals with the prerequisites of creating content in many languages, incl. technologies and standards related to character encoding, language identification, font selection etc. Internationalization is a prerequisite of *localization*; the adaptation of content to local markets and cultures, which typically involves *translation* and is often outsourced to Language Service Providers (LSPs). Finally, ever growing volumes of content and numbers of target languages make use of *language technologies* (e.g. machine translation) a key to achieving a multilingual Web. Attending communities of the major conference series in their respective areas, i.e. Localization World, LREC (Language Resources Evaluation Conference), and the Unicode conference are largely disjoint. Thus, a major success of the MultilingualWeb was even bringing together important stakeholders from the disjoint areas, and this has been successfully perpetuated in the formation of the MultilingualWeb-LT WG.

3. MULTIPLE VIEWPOINTS
Highlighting and reconciling the differing viewpoints of the concerned communities was a major challenge for the MultilingualWeb workshop series.

3.1 Developers
Platform Developers provide the building blocks that are needed for multilingual content creation and access on the Web. Many of these technologies are still rapidly evolving and web browsers play a crucial role. Speakers addressed the enhancement of character and font support, locale data formats, internationalized domain names and typographic support. One major gap in this area comprises tackling of translation workflows. Although more web content is being translated, the key web technology HTML so far has no means to support this process, and was identified as a priority for the W3C and browser implementers. Another gap is the range of content formats and technology stacks in use. While HTML5 plays a crucial role in the future of web content development, its relation to other forms of digital content has not become clear yet, e.g. in relation of multi-media content and XML-based, component-oriented documentation.

3.2 Creators
Creators need to bring content to different delivery platforms, especially mobile devices. Since these devices lack computing power, many aspects of multilinguality need to be carefully addressed. Content creation must also support voice-based and multimodal applications, or short messages delivered by social media or SMS. Navigation of web content across languages is

another area that lacks standardized approaches and best practices. In all cases content creators need standardized ways to identify non-translatable content and other instructions downstream to localization processes.

3.3 Localizers

Localizers deal with internationalization practice in content creation, the distribution of content to LSPs and the onward distribution to individual translators. Workflow automation for improved efficiency of this process requires improved standards for metadata that accompany content throughout the cycle. While the complex and fast changing nature of content itself presents a challenge, so does the fragmentation of standardization efforts in this area. Multiple, sometime overlapping, standards are available from different international organizations including the W3C, the International Organization for Standardization (ISO), Organization for the Advancement of Structured Information Standards (OASIS), European Telecommunications Standards Institute (ETSI), the Unicode Consortium and the now defunct Localization Industry Standards Association (LISA). The issue here is often just to understand how the standards interplay.

3.4 Machines

For machines, i.e. applications based on language technology, the need for standardization related to metadata and the localization process is of utmost importance. Language resources are crucial here for the training of data-driven language technologies, including their standardized representation and means to share resources. It became clear that machine translation technology developers, creators, and localizers need to work tighter together for better translation quality. Without at least partially automated translation, valuable web resources like Wikipedia will continue to be available to only a small proportion of the global population.

3.5 Users

It became clear that the worldwide end user interest in multilingual content is high, but significant organizational and technical challenges exist to integrate less developed economies in linguistically diverse regions of Asia and Africa.

Multilingual social media are becoming more important and can be supported by on-the-fly MT. However it is important to have a clear border between controlled and uncontrolled environments of content creation and translation. Thus, high quality translation can be differentiated from automated results suitable only for gisting.

3.6 Policy Makers

Many gaps related to the multilingual web are not technical, but are related to e.g. political decisions about the adoption of standards. In the localization and language technology area, proprietary solutions prevailed for a long time. The vision of an open Web available in all languages requires a radical change, and MultilingualWeb will play a crucial role in bringing the right people together to tackle not only the technical foundations, but also to convince the relevant decision makers, and to address commercial concerns, e.g. through appropriate licensing standards for language resources.

4. INTEROPERABILITY LANDSCAPE FOR THE MULTILINGUAL WEB

The MultilingualWeb workshop series paints a compelling picture of the direction being taken by Web stakeholders in embracing its multilingual future. Organizations increasingly use the Web as their primary means of communicating with customers and stakeholders. An organization's web content is continuously generated by a large range of internal and external users, requiring *Content Management Systems* to ensure it is maintained in a coherent and navigable state. To target international markets or multilingual national groups, organization must also *localize web content*, so that it can be presented effectively across many languages and cultures. Localization is typically outsourced to Language Service Providers (LSPs) who employ translators supported by *language technologies* such as machine translation (MT) and translation memory to deliver localized content to tight cost, time and quality constraints. Organizations may also decide to directly apply MT to web content in situations where its volume or transient nature preclude the expense of quality-assured translations offered by LSPs. Users can also avail of language technologies directly with open web services, e.g. translation services from Google and Microsoft, or OpenCalais text analytics service from Thomson Reuters.

However, the smooth interoperation between diverse, changing web content and both localization workflows and language technologies remains a challenge. Up to 20% of localization costs can be attributed to manual content transform overheads [1].

To support the interoperability requirements analysis, we identify the main distinct functional areas evident in industry, characterized by either distinct market sectors, by specific human functions, or by classes of supporting technologies. These areas, depicted in figure 1, can be defined as follows:

Content Management from the 'creators' topic; the associated human functions are concerned with content generation and consumption. This function is supported by Content Management Systems (CMS) that support web content either as distinct web pages, or documents, or via content components that are flexibly composed on-demand to suit different consumption requirements.

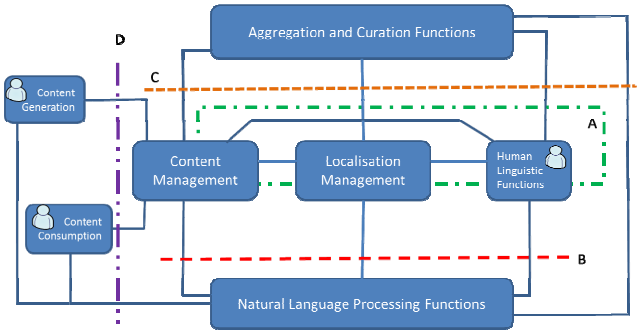

Figure 1: An interoperability map for the Multilingual Web

Localization Management from the 'localizers' topic deals with the industrial process of translating and adapting content to the languages and cultural norms of different locales. This function is supported by specialized workflow management systems, called Translation Management Systems (TMS), translation memories (TM) (databases of previous translations designed to maximize translation reuse) and terminology management systems. The associated human linguistic functions include translation, post-editing (correcting of deficiencies in MT), source and target language quality assurance, and cross lingual terminology management. System support is typically offered to these workers through Computer Aided Translation (CAT) tools.

Natural Language Processing (NLP) from the 'machine' topic represents the maturing class of language technologies that are

increasingly relevant for the multilingual web. These technologies are important for localization industry, looking to improve its throughput while facing a ballooning demand driven by globalization, downward price pressure per translated word and the limited number of professional translators available. Principle is machine translation. Recently, data driven techniques, and in particular statistical machine translation (SMT), have led to a resurgence of interest in MT. Provided suitable volumes of (clean) parallel text are available, SMT can be comparatively quickly and relatively cheaply applied to new domains and language pairs. Combinations of language knowledge encodings and data driven approaches can offer other NLP solutions relevant to the multilingual web. Text analytics can support named entity recognition for terminology management, but is also finding use in sentiment analysis of social media. A further class of NLP potentially important for Multilingual Web is speech processing.

Aggregation and Curation includes the collection, classification, indexing and searching of web content in large volumes. This can be both monolingual, with Web search being the primary application here, and multilingual, i.e. cross lingual search and search engine optimization. The NLP Research and Development community is also active in the collection and assembly of language resources, a term used to refer to a wide range of language corpora, including parallel text, transcribed speech audio, semantic annotations of mono-lingual content etc. To date, however, the integration of such language resources with the use of NLP in content and localization management remains limited. Language resources curation plays some role in support of localization, as collecting and curating translation memories and term bases for future jobs.

Within this framework, the *current landscape of interoperability standards* can be summarized as follows:

Content and Localization Management Systems form the backbone of the current multilingual content processing industry. It is supported by established technology markets which are moving to better integration through standards such as: XLIFF (XML Localization Interchange File Format)[2]; TMX (Translation Memory Exchange)[3]; TBX (Term Base Exchange)[4]; SRX (Segmentation Rules Exchange)[5]; and ITS (Internationalization Tag Set)[6]. There are also established XML and emerging RDF (Resource Description Framework) formats for content, lexical and metadata serialization, including: DITA (Darwin Information Typing Architecture – XML component content management standard)[7]; linguistic annotation [8]; and lexicon model for ontologies[9]. There are also standard APIs for manipulating content, e.g. DOM (Document Object Model), CMIS (Content Management Interoperability Services)[10]. These address interoperability points within zone A in figure 1.

Natural Language Processing, as emerging set of technologies, is being adopted on a more ad hoc basis. Integration typically happens via simple content transform or annotation web services. However, these lack common semantics for measuring and assessing quality, reliability, staleness, etc. across different service offerings (interoperability spanning boundary B in figure 1).

Aggregation and Curation functions must deal with collecting and adding value to large collections of data from multiple sources, including web pages/documents, search applications or social media streams for sentiment analysis applications. Data and metadata interoperability are still largely siloed by media and application type and by proprietary document/media formats, or XML vocabularies, complicating interlinking/analysis across sources. These functions are also key to providing low-cost, application-specific language resources as training corpora for NLP functions. Often, existing language resource exchange formats, such as TMX and TBX are used, but these lack metadata fields that may be relevant for training purposes. Interoperability points span boundary C in figure 1.

Human-Content Interaction is largely web-based, increasingly delivered over multiple media/modalities via mobile and embedded devices. Though the interoperability of the presentation of diverse content is being addressed by developments in HTML5, the portability of user interaction preferences for adaptation and personalization across sites, data sources, devices and media remains a challenge, as well gathering explicit quality feedback and business intelligence from users. These interoperability points span boundary D in figure 1.

5. WEB METADATA FOR LANGUAGE RELATED TECHNOLOGY IN THE WEB

Above, we raised a number of big standardization challenges, yet effective progress must be made in small achievable steps. Thus, we identified that standardized definition of metadata related to the multilingual characteristics of web content could have high impact in resolving some of the above interoperability issues with minimal disturbance to existing technologies. Such an approach had already been successful in W3C ITS [6] that defined a small set of independent data categories to be used for XML-based (web) content annotation at different levels of granularity. Each category conveys information about a multilingual characteristic of the annotated content, e.g. whether it should be translated or not, whether it represents a defined term, information about reading direction, or language specific annotation rendering.

This approach is now extended to address gaps in standards efforts between content management, localization workflow, and language technologies. Three use cases have been prioritized:

CMS – LT Interaction: The direct interaction of language technologies with content management systems. An example would be identifying to machine translation services, which text on a web page should not be translated or which should be translated as specific terminology. Another example is the use of text analytics services to annotate some text, e.g. as a named entity, i.e. a candidate for terminology management.

CMS – Localization Roundtrip: Support for reliable transmission of internationalization metadata from content creation to localization. This also requires that web content can maintain localization related metadata for access by multiple localization functions, e.g. translation and review.

Content and Localization Management – Language Resources: Ensuring that metadata related to internationalization and quality of translation process can be maintained with the content for later ascertaining its suitability as NLP training corpora, e.g. bi-text for SMT training, or for cross-lingual information retrieval.

These metadata standardization requirements are now being addressed by the Multilingual Web W3C working group (http://www.w3.org/International/multilingualweb/lt/).

5.1 MultilingualWeb-LT Working Group

The MultilingualWeb-LT project and WG will address the integration gaps between content management systems operated by content owners, the localization management and workflows at LSPs, and the emerging role of LT. The core EC funded consortium consists of 13 partners from the localization industry, academe, as

well as content owners and creators, and has established within W3C the open MultilingualWeb-LT WG:

Microsoft represents large localization service clients, content creators, owners, publishers, and social media stakeholders. Cocomore offers CMS based solutions to meet their client's managed web presence needs. Moravia Worldwide, ENLASO, Linguaserv and VistTEC represent the needs and views of LSPs. Lucy Software represents LT vendors (as also some of the LSPs) while Dublin City University (DCU) offers expertise in SMT and Jožef Stefan Institute (JSI) in text analytics. German Research Centre for Artificial Intelligence (DFKI), Trinity College Dublin (TCD), University of Economics Prague (VŠE Praha), and the University of Limerick (UL) bring their experience in open standards creation and next generation localization research.

Apart from the above described group, 8 more W3C members have joined the WG during the charter review process. Together, the WG members have planned 7 product level reference implementations, 11 metadata implementations in their operations, and 8 experimental platform implementations. It is highly desirable that other W3C members continue joining the group in order to strengthen the representativeness of consensus and to expand the number of planned implementations.

The WG charter sets four main goals: 1) To develop the successor of ITS 1.0 (ITS 2.0). 2) To concentrate on the use of these data categories in HTML5 and CMS or XML from which HTML pages are generated. 3) To formally and consistently define processing requirements of data categories. 4) To foster reference implementations of the data categories in Web-related environments such CMS, TMS, MT systems etc.

Other categories will be considered in the areas of translation provenance (human and machine translation of different types); human or automated post- and pre-editing, including degree of post-editing; legal metadata pertaining to ownership and usage rights; Quality Assurance (QA);topic or domain information etc. The project will build several, mostly open source, reference implementations in the following three priority areas:

Online MT Systems will be made aware of the metadata, enabling more adequate translation results, and sensitive to the outputs of the modified CMS described above.

Integration of CMS and Localization Chain (TMS and bitext management in general). Open source modules for the Drupal CMS will be built with support of the metadata that will then be taken up in web-based tools to support the localization chain: from the process of gathering of localizable content, the distribution to translators, to the re-aggregation of the results into localized output. The open standard bitext format XLIFF will play a key role in localization round-tripping of the metadata.

MT Training. Metadata aware tools for training MT systems will be built. Again these are closely related to CMS that produce the necessary metadata for quality training corpora.

6. RELATED WORK

Other EC-funded projects play an important role in relevant technology, community, and consensus building:

FLaReNet (Fostering Language Resources Network - www.flarenet.eu) with their "Blueprint of Actions and Infrastructures", which is a set of (technical) infrastructure, R&D, and politics related recommendations [11].

META-NET (www.meta-net.eu) is dedicated to fostering the technological foundations of a multilingual European information society, by building a shared vision and strategic research agenda, an open distributed facility for the sharing and exchange of resources (META-SHARE, www.meta-share.eu, addressing intellectual property issues), and by building bridges to relevant neighbouring technology fields.

PANACEA (www.panacea-lr.eu) involves wrapping open source or partner-provided software functions as web services and integrating them into data process workflows that are constructed, run and monitored using the TAVERNA open source service composition tool for scientific data processing workflows.

LetsMT! (www.letsmt.eu) supports the important enabling open source project *M4Loc* (http://code.google.com/p/m4loc/), which is bringing critical localization standards support to the plain text based Moses SMT toolkit. While proprietary and closed source inline tagging support exists for Moses in commercial offerings, M4Loc brings an open-open solution. MultilingualWeb-LT reference implementations will make use of M4Loc and will enhance it with newly developed metadata categories.

MONNET (www.monnet-project.eu) focuses on localization of ontologies and thereby enabling multilingual access to organizational knowledge management activities, incl. corporate accounting governmental information access for public.

These and the MultilingualWeb itself show that a holistic view of a truly multilingual web is emerging, overcoming gaps between internationalization, localization, and language technology.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the leadership and vision of Richard Ishida in coordinating the MultilingualWeb under W3C Internationalization activity. This paper has been supported by EC via the MultilingualWeb-LT project (contract number 287815) and by the Science Foundation Ireland (Grant 07/CE/I1142) via the Centre for Next Generation Localisation (www.cngl.ie) at UL and TCD.

8. REFERENCES

[1] Lack of Interoperability costs the translation industry a fortune, "Report on a TAUS/LISA survey on translation interoperability", 25 Feb, 2011, TAUS

[2] XLIFF Version 1.2, OASIS Standard, 1 February 2008

[3] Translation Memory Exchange (TMX), http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

[4] Term Base Exchange (TBX), http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf

[5] Segmentation Rules Exchange (SRX), http://www.gala-global.org/oscarStandards/srx/srx20.html

[6] Internationalization Tag Set (ITS) Version 1.0, W3C Recommendation 03 April 2007

[7] Darwin Information Typing Architecture (DITA) Version 1.2, OASIS Standard, 1 December 2010

[8] Ide, N., Romary, L. (2004) International standard for a linguistic annotation framework, Journal Natural Language Engineering, Vol 10 Iss 3-4, September 2004

[9] Declerck T., et al, (2010) lemon: An Ontology-Lexicon model for the Multilingual Semantic Web, W3C Workshop, Madrid 2010

[10] Content Management Interoperability Services (CMIS) Version 1.0, OASIS Standard, 1 May 2010

[11] FLaReNet "Blueprint of Actions and Infrastructures", 15 Dec 2010, http://www.flarenet.eu/sites/default/files/D8.2b.pdf