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
In this paper we report in the Multilingual Web initiative which is a collaboration between the Internationalization Activity of the W3C and the European Commission (EC), realized as a series of EC-funded projects. We review the outcomes of the first of these projects, "Multilingual Web", which conducted a successful workshop series aimed at analyzing borders or "gaps" within Web technology standardization that currently hinder multilinguality on the Web. The resulting insight into the scientific, industrial and user stakeholders involved led to a further project "MultilingualWeb-LT". This project has established a cross-sector W3C Working Group that will address some of the gaps that have been identified through standardization of meta-data.

Categories and Subject Descriptors
H.5.4 Hypertext/Hypermedia

General Terms
Standardization

Keywords
Language Technologies, Localization, Internationalization, Web Technologies, Standardization, metadata, interoperability

1. MULTILINGUAL WEB: OVERVIEW
MultilingualWeb Initiative (http://www.multilingualweb.eu/) began as a thematic network project funded by European Commission (EC), exploring standards and best practices that support the creation, localization and use of multilingual web-based information. It is lead by the World Wide Web Consortium (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). The project held a series of four public workshops held between October 2010 and March 2012 in Madrid, Pisa, Limerick and Luxembourg respectively. Their focus was the standards and best practices currently exist to enable a fully multilingual web, and what gaps still needed to be filled. They have been of enormous success, in terms of the number of participants, awareness via social media, and the outcome of discussions.

Formation of the W3C Working Group (WG) MultilingualWeb – Language Technology (MultilingualWeb-LT), on 7th March 2012 arose directly from this success. This WG, again with the financial support of the EC, will develop meta-data standards for Web content that facilitate its seamless interaction with language technologies to enable multilingualism and localization processes. This will be achieved with broad industry consensus under the W3C Internationalization Activity. MultilingualWeb-LT Charter has been endorsed by 24 W3C members (11 of them outside the EC funded core group) including Adobe, BBC, Boing, Microsoft, Opera and representatives of the Indian government.

This paper reviews the outcomes of the initial MultilingualWeb workshop series and describes the scope of the MultilingualWeb-LT standardization plans.

WWW'12, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/11/04.

2. STAKEHOLDER COMMUNITIES
The MultilingualWeb initiative recognized the need to bridge between several distinct communities with different vital roles to play in achieving widespread multilingualism on the web. The web content management industry sees the problem as one of *Internationalization. Internationalization* deals with the prerequisites to create content in many languages. This involves technologies and standards related to character encoding, language identification, font selection etc. The proper internationalization of web content and technologies is required for its *localization*. Localization is the adaptation of content to local markets and cultures, which typically involves *translation* and is often outsourced to Language Service Providers (LSPs). Finally, with ever growing volumes of content in need of translation, and a growing number of target languages, the use of *language technologies* (e.g. machine translation) will be key to achieving a multilingual Web.

The lack of integration between these communities is demonstrated by the disjointedness of their attendance at the major conferences the areas, such as Localization World, LREC (Language Resources Evaluation Conference), and the Unicode conference series. A major success of the MultilingualWeb project, therefore, was bringing together important stakeholders from the areas of internationalization, localization and language technologies and this has been successfully perpetuated in the formation of the MultilingualWeb-LT WG.

3. MULTIPLE VIEWPOINTS
Highlighting the differing viewpoint of the concerned communities was a major challenge for the MultilingualWeb workshop series. In order to identify the gaps that existed discussion were structured according to the following topics.

3.1 Developers
Platform Developers provide the technological building blocks that are needed for multilingual content creation and access on the Web. Many of these building blocks are still rapidly evolving and

web browsers play a crucial role. In the workshops, speakers addressed the enhancement of character and font support, locale data formats, internationalized domain names and typographic support. One major gaps in this area is related to handling of translation worflows. 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 more and more need to bring content to different delivery platforms, especially via mobile devices. Since these devices lack computing power, many aspects of multilinguality (e.g. usage of large fonts) need to be carefully addressed. Content creation must also support voice-based and multimodal applications, or short messages delivered by social media or SMS, challenging traditional text-based internationalization and localization techniques. Navigation of web content across languages is another area that lacks standardized approaches and best practices. In all cases content creators need standardized way to identify non-translatable content and other translation instructions downstream to localsiation 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. Improved efficiency of this process requires between technical integration through improved standards for the meta-data that accompanies content through the resulting workflow. 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 Organszation for Standardisation (ISO), Organization for the Advancement of Structured Information Standards (OASIS), European Telecommunications Standards Institute (ESTI), the Unicode consortium and the now defunct Localization Industry Standards Association (LISA). The gap 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 outmost 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 closer integration of machine translation technologies developer, creators and localizers is a major requirement for the better translation quality. Machine translations are also crucial bridging between languages already well supported by the language industry and those less so. 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

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 approached for reaching people in less developed economies, especially in linguistically divcerse regions such in Asia and Africa.

Multilingual social media are becoming more important and can be supported by language technology via on-the-fly machine translation. However it is important to have a clear border between controlled and uncontrolled environments of content creation and translation. Only in this way can high quality translation be differentiated from automated results suitable only for ‘gisting’.

3.6 Policy Makers

The topic of policy makers it is of high importance: many gaps related to the multilingual web are not technical ones, but are related to e.g. political decisions about the adoption of standards. For example, in the localization and language technology area, proprietary solutions prevailed for a long time. The vision on an open Web available in all languages requires a radical change, and MultilingualWeb will play a crucial role in bringing the right people together in both constructing the technical foundations and for convincing the relevant decision makers and for making sure commercial concerns are addressed, e.g. through appropriate licensing standards for language resources.

4. INTEROPERABILITY LANFSCAPE FOR THE MULTILINGUAL WEB

The Multilingual Web workshop series paints a compelling picture of the direction being taken by Web participants in embracing a multilingual future. Clearly, 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 and translation memory to deliver localized content to tight cost, time and quality constraints. Organizations may also decide to directly apply machine translation to web content in situations where its volume or transient nature precludes the expense of the 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 and the 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. One recent industry survey with up to 20% of localization costs being attributed to manual content transform overheads [1].

Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged. It can be addressed in terms of

Functional Interoperability, resulting in shared architectures, methods and frameworks, and of Semantic Interoperability, resulting in shared data types and terminology/coding.

To support the analysis of the interoperability requirements for the multilingual web, it is informative to identify the main distinct functional areas evident in industry today, characterized by either distinct market sectors, by specific human functions or by classes of supporting technology. These areas, depicted in figure 1, can be defined as follow:

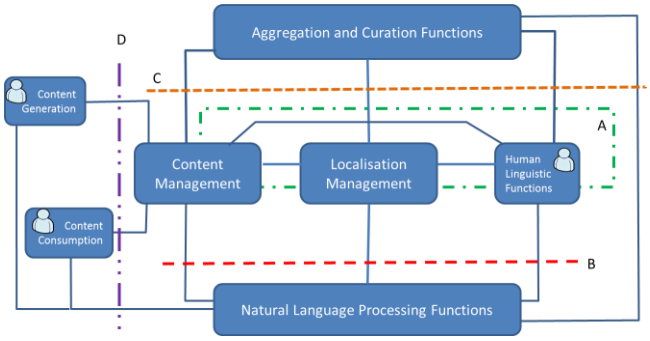

Figure 1: An interoperability map for the Multilingual Web

Content Management: Discussed under the 'creators' topic, this deals with the lifecycle of content from generation to consumption. The associated human function 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.

Localization Management: Discussed under the 'localizers' topics, this 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, sometimes 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 definciencies in automated translation), 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): Discussed under the 'machine' topic, NLP represents the maturing class of language technologies that are being brought to bear on the Multilingual Web. These technologies are of significant interest to industry looking to improve the throughput of the localization industry in the face of ballooning demand driven by globalization, downward price pressure per translated word and the limited number of professional translators available. Principle amongst these is machine translation. Rule-based machine translation is a mature technology that can yield good results but only after years of effort in encoding linguistic knowledge. This therefore limits its expansion to new content domains and language pairs. More recently, data driven techniques, and in particular statistical machine translation (SMT), have led to a resurgence of interest in MT. Provided suitable volumes of (reasonably clean) parallel text (translated sentences ordered as bitext, i.e. as a sequence of source-target pairs) 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 in support of terminology management, but is also finding use in sentiment analysis of social media on the web. A further class of NLP potentially important s for Multilingual Web is speech processing, with speech recognition and speech synthesis being increasingly integrated into web applications. The automation of multilingual speech processing is still a major challenge however.

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. Language resource curation plays a major role in support of localization, with the collection of sentence and term translations being curated for use in future jobs as translation memories and term bases respectively. 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.

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

Content and Localization Management Systems. These form to 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 meta-data serialization, including: DITA (Darwin Information Typing Architecture – an XML based 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 is 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: These must deal with collecting and adding value to large collections of data from multiple different sources, including web pages/documents for search applications or social media streams for sentiment analysis applications. Data and meta-data 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 meta-data fields that may be relevant for training purposes. Interoperability points span boundary C in figure 1.

Human-Content Interaction: This is largely web-based, but increasingly delivered over multiple media/modalities via mobile and embedded devices. Though the interoperability of the presentation of diverse content is being increasingly 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. Equally importantly, gathering explicit quality feedback and business intelligence from users remains a challenge. These interoperability points span boundary D in figure 1.

5. WEB METADATA FOR LANGUAGE RELATED TECHNOLOGY IN THE WEB

While the landscape view offered in the previous section raises a number of major standardization challenges, effective progress must be made in small achievable steps. To this end we identified that standardized definition of meta-data related to the multilingual characteristics of web content could offer high impact in resolving some of the above interoperability issues with minimal disturbance to existing technologies. Such an approach had already been successfully demonstrated in the ITS standard [6]. This defined a small set of independent data categories could be used to annotate web content at different levels of granularity. Each data category conveyed information about the multilingual characteristic of the annotated content, e.g. whether it should be translated or not, whether it represented a defined term, information about reading direction or language specific annotation rendering, i.e. the Ruby notation used in Japan.

This approach is therefore extended to address the current disjoint in standards efforts between content management, localization workflow and language technologies. Specifically three use cases have been prioritized for attention by standardized meta-data:

1. **Content Management - Language Technology Interaction**: The direct interaction of language technologies with content management systems. Exmaple would be indentifying to machine translation services which text on a web page that should not be translated or which should be translated as specific terminology. Another example is the use of text analytics services to annotate certain text, e.g. as a named entity that is a candidate for terminology management.
2. **Content Management – Localization Roundtrip**: Supporting the reliable transmission of internationalization meta-data from content creation to localization functions. This also requires that web content can maintain localsiation related meta-data when access via multiple loclsaiiton functions, e.g. translation and translation review
3. **Content and Localization Management - Language Resource**: Ensuring that meta-data related to internationalization and the quality of the translation process can be maintained with the content for later ascertaining its suitability as NLP training corpora, e.g. bi-text for SML training, or as the subject of cross-lingual information retrieval.

These meta-data 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 that exist between content management systems operated by localization clients as content owners, the localization management and workflow services by LSPs, and the emerging role of language technologies. The core EC funded consortium consists of partners from the localization industry, academe, as well as content owners and creators.

Overall thirteen international partners are collaborating in the core consortium that has established within W3C the open MultilingualWeb-LT WG:

- Microsoft represent the viewpoint of large localization service clients, content creator, owner, and publisher, and social media stakeholder.
- 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 Language Service Providers.
- Lucy Software represents Language Technology vendors (a role also played by some of the LSPs) while Dublin City University offer expertise in statistical machine translation and Jožef Stefan Institute (JSI) in text analytics platforms
- 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 core group, 8 more W3C members have joined the WG during the charter review process, namely: Adobe, BBC, Consiglio Nazionale delle Ricerche, Department of Information Technology, Government of India, DERI Galway at the National University of Ireland, Ecole Mohammadia d'Ingenieurs Rabat (EMI) Universidad Politécnica de Madrid and Vrije Universiteit. Together, these WG members have so far planned seven product level reference implementations, eleven plan to use the metadata in their operations and eight plan to develop experimental platforms making use of the standardized metadata. It is highly desirable that other W3C members continue joining the group in order to strengthen the representativeness of consensus and expand the number of planned implementations.

The WG charter specifies: The MultilingualWeb-LT Working Group has four goals:

1. To develop the successor of ITS 1.0.
2. To concentrate on the use of these data categories in HTML5 and "deep Web" content, for example a CMS or XML files from which HTML pages are generated. This does not exclude the definition of additional data categories (see below), but describes the focus of the Working Group.
3. To define processing requirements of data categories formally and in a consistent manner.
4. To foster reference implementations of the data categories in other, XML and non-XML environments that are on the Web or closely related to the Web: CMS systems, Web based

localization chain services, online machine translation systems.

In addition it will consider data categories 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) provenance including application of translation QA, results of QA and human and tool input into QA assessment; and Topic or domain information

The project will establish several, mostly open source, reference implementations around the three priority areas, in which metadata is being used:

- **Online MT Systems**. MT systems will be made aware of the metadata, which will lead to more satisfactory translation results. An online MT system will be made 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 that support the creation of the metadata. The metadata will then be taken up in web-based tools that 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. They will lead to better quality for MT training corpora harvested on the Web.

6. RELATED WORK

Other EC-funded projects play an important role in community and concensus building and in open technology demonstrators in the area of language technology interoperability research and industrial applications. Key projects are:

- *FLaReNet* (Fostering Language Resources Network - www.flarenet.eu) has developed a common vision for the area of language resources. The FLaReNet "Blueprint of Actions and Infrastructures" is a set of recommendations to support this vision in terms of (technical) infrastructure, R&D, and politics [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 – which address much needed intellectual property issues), and by building bridges to relevant neighbouring technology fields.
- *PANACEA* (www.panacea-lr.eu) addresses the use of web service and service composition to ease the integration of language technology functions. The approach involves wrapping open source or partner-provided software functions as web services and integrating them into data processes workflows that are constructed, run and monitored using the TAVERNA open source service composition tool for scientific data processing workflows.
- *LetsMT!* (www.letsmt.eu) is trialing a sharing repository for SMT training resources. This may offer insight from its experience in corpora sharing and use of existing public corpora and to establish a migration path from their centralized sharing and static data model, to future, decentralized web-based sharing and extensible data models for training corpora. LetsMT! supports the important 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 exist for Moses in commercial offering, M4Loc brings an open, robust standards based (XLIFF and TMX) solution under LGPL license. MultilingualWeb-LT reference implementations will make use of M4Loc tools and will enhance them with newly developed metadata categories.
- *MONNET* (www.monnet-project.eu)focuses on the localization of ontologies and thereby on enabling multilingual access to the output of organizational knowledge management activities, ranging from multilingual access to corporate accounting information to citizen access to governmental information in multiple languages.

These demonstrate that a holistic view is emerging, in which the differences between internationalization, localization and language technology are being bridged towards the aim of a truly multilingual web.

7. ACKNOWLEDGMENTS

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

8. ACKNOWLEDGMENTS

[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