D6.2.2: SUMMARY REPORT 2

Arle Lommel (DFKI), Felix Sasaki (DFKI)

Distribution: Public

MultilingualWeb-LT (LT-Web)
Language Technology in the Web

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	6.2.2
Deliverable title:	Summary Report 2
Dissemination level:	PU
Contractual date of delivery:	31st May 2013
Actual date of delivery:	31st May 2013
Author(s):	Arle Lommel, Felix Sasaki
Participants:	DFKI
Internal Reviewer:	-
Workpackage:	WP6
Task Responsible:	Arle Lommel
Workpackage Leader:	Arle Lommel

The workshop report has been published online as HTML at

http://www.multilingualweb.eu/documents/rome-workshop/rome-workshop-report

This document contains a PDF version of the HTML page.

Rome report

W3C Workshop Report: The Multilingual Web – Making the Multilingual Web Work

12 - 13 March 2013, Rome

Today, the World Wide Web is fundamental to communication in all walks of life. As the share of English web pages decreases and that of other languages increases, it is vitally important to ensure the multilingual success of the World Wide Web.

The MultilingualWeb initiative examines best practices and standards related to all aspects of creating, localizing, and deploying the Web multilingually. The initiative aims to raise the visibility of existing best practices and standards and to identify gaps in web technologies that impact multilinguality online. The core vehicle for this effort is a series of events that started in 2010 (run by the initial MultilingualWeb Project and now by its successor, the MultilingualWeb-LT project).

On 12–13 March 2013 the W3C ran the sixth workshop in the series in Rome with the theme entitled “Making the Multilingual Web Work.” The Rome workshop was hosted by the Food and Agriculture Organization (FAO) of the United Nations and a little over 150 people attended it. Daniel Gustafson, Deputy director general of FAO, gave a brief welcome address.

As with the previous workshops, this workshop focused on discussion of best practices and standards aimed at helping content creators, localizers, tools developers, and others meet the challenges of the multilingual Web. The key objective was to bring together speakers and to provide opportunity for networking across a wide range of communities.

The Workshop featured one and a half days of talks plus a final half day dedicated to an Open Space discussion forum held in breakout sessions. In the Open Space portion participants dynamically suggested ideas for discussion groups and then split into six groups. Each of these groups reported back in a plenary session at the end of the session. Participants could join whichever group they found interesting, and could switch groups at any point. During the reporting session participants could ask questions of or make comments about the findings of other groups, providing a way for them to learn about the results of all of the groups. This, once more, proved to be a popular part of the workshop.

We were able to stream the content live on the Internet and we also recorded the presentations, which will be made available on the Web using the VideoLectures service. We also once more made available live IRC scribing to help people follow the workshop remotely and so assist participants in the workshop itself. Attendees were encouraged to Tweet about the conference and the speakers during the event, and you can see these linked from the program page.

The program and attendees continued to reflect the same wide range of interests and subject areas as in previous workshops and we once again had good representation from content creators and the localization industry as well as research and the government/non-profit sector.

After a short summary of key highlights and recommendations, this document provides a short summary of each talk accompanied by a selection of key messages in bulleted list form. Links are also provided to the IRC transcript (taken by scribes during the meeting), video

recordings of the talk (coming soon), and the talk slides. Most talks lasted 15 minutes, although some speakers were given longer slots. Finally, there are summaries of the breakout session findings.

Summary

What follows is an analysis and synthesis of ideas brought out during the workshop. It is very high level, and you should watch or follow the individual speakers talks to get a better understanding of the points made.

Our keynote speakers, Mark Davis and Vladimir Weinstein, talked about the ways in which Google is working to address internationalization and localization challenges that go beyond translation: issues like dealing with address, name, and telephone number formats, working with languages with limited font support, and ensuring consistent international support across a large portfolio of products. Because these challenges are faced by many companies and organizations, shared solutions provide the best return and Google is actively contributing its solutions into the open-source community.

In the Developers session we heard from Jan Anders Nelson about the importance of apps on Windows8 and Microsoft's work to simplify localization tasks. He provided a demo of tools that Microsoft provides to app developers that support localization into many languages and that allow cross-platform (e.g., desktop and mobile) leveraging.

Workshop sponsors

Gavin Brelstaff and Francesca Chessa provided an update on work discussed at the MultilingualWeb Workshop in Pisa on using standards-based approaches to provide audio-text synchronization in online video. Their solution provides easy-to-use tools that allow users to provide synchronized captions in multiple languages, with both line- and word-level alignment.

Gábor Hojtsy described the ways that Drupal supports localization of both the core application and modules, as well as content. By using a common repository of translations, maximum reuse is encouraged and translators can contribute as much or as little as they desire. Currently the project is working on improvements for localization of user-editable content.

Reinhard Schäler discussed the requirements for "non-market localization" (the language requirements needed to reach the world's citizens, regardless of whether they are customers). The goal of reaching the broader world requires new workflows that simplify contributions and enable people to contribute without a heavy infrastructure. To further this effort, the Rosetta Foundation is launching the "Trommons" (Translation Commons), where translators can contribute their work for reuse.

During the Creators session, Román Díez González and Pedro L. Díez-Orzas discussed the ways in which the forthcoming Internationalization Tag Set (ITS) 2.0 standard improved localization process for the Spanish Tax Agency. In particular, they discussed how Real Time Translation Systems (RTTS) have tended to cause problems by taking needed control away from content creators, but combining HTML5 with ITS 2.0 returns control over terminology, domain, and other key features while simultaneously reducing costs and overhead. Hans-Ulrich von Freyberg then spoke about the German Industrial Machine Builders' Association (VDMA)'s experience with ITS 2.0 and Drupal. By using ITS 2.0, the VDMA has been able to reduce delivery time for translated texts by 30% while improving quality through the use of automated content-enrichment tools. The savings make it possible for VDMA to expand its volume of translated content.

Brian Teeman spoke about the ways in which Joomla, which is used for 2.8% of sites on the Internet, is working to simplify localization for sites deployed on multiple output devices (desktop, mobile, etc.). Joomla is also working with community-based translation services to help provide options for site creators that want to localize their content.

Vivien Petras spoke about the semantic and interoperability challenges faced in building Europeana, a pan-European repository of cultural materials. Currently most metadata is monolingual and most users come into the repository through Internet search results. As a result, Europeana is working to provide an enriched set of metadata based on linked open data principles that can cross language boundaries.

Christian Lieske and Inna Nickel presented on the importance of language quality checking to improve translation quality. They discussed use of LanguageTool, an open-source language quality-checking tool, to meet these needs and the ways in which integration with ITS 2.0 and various tools makes it an attractive option for addressing language quality.

The Localizers session began with a presentation from Bryan Schnabel about how integration of various CMS and localization components with the XLIFF standard has enabled Textronix to simplify its localization processes and reduce costs by eliminating significant amounts of previously manual work. XLIFF also simplifies working with DITA content since it can bundle what previously would have been hundreds of topic-based files into single XLIFF files, thus simplifying the process of translation.

Sinclair Morgan presented results of studies on the effectiveness of machine translation and post-editing (MT+post-editing) for high-quality translation. When properly applied to narrow domains with good linguistic resources in an ergonomic environment, MT+post editing can reduce costs and time to delivery (~30%) significantly over traditional manual translation methods.

Charles McCathie Nevile discussed the difficulties faced by large companies in meeting their localization requirements. While there are technology solutions that are needed, companies also need to be aware of their internal biases towards particular kinds of solutions or approaches and need to work to overcome them if they are to succeed.

The session ended with a talk from Hans Uszkoreit about current projects to address barriers to translation quality on the Internet. He provided an overview of the QTLaunchPad and META-NET projects and their approaches to providing resources to assist European entities in achieving better quality. By removing barriers greater participation in vital discussions will be possible for European citizens. Machine Translation will have to play a part in meeting these needs, so current efforts are focused on improving MT quality in meaningful ways.

The second day began with Machines session. José Emilio Labra Gayo provided an overview of best practices for creating multilingual linked open data (LOD) patterns. He highlighted the difficulties faced in localizing LOD patterns, and then provided solutions with examples and a brief discussion of those solutions. After José finished, Asunción Gómez-Pérez's presentation highlighted similar issues, with a focus on the architectural and process requirements needed to ensure multilinguality in linked open data. She also highlighted the ways in which multiple services must interact to ensure multilinguality.

Peter Schmitz discussed the EU Publication Office's CELLAR repository, which provides metadata about official EU publications. The Office also provides multilingual controlled vocabularies and has contributed to the European Legislation Identified (ELI) standard.

Gordon Dunsire presented on the International Federation of Library Associations and Institutions (IFLA) and its standards for multilingual bibliographic and library information. He discussed issues faced in this area and IFLA's moves to codify best practice.

Thierry Declerck discussed joint work with Max Silberztein on the NooJ linguistic development environment, which allows users to automatically annotate text with important linguistic information.

The Users session began with a talk from Pat Kane on Internationalized Domain Names (IDNs). He discussed the issues of trust and security that impact adoption. In addition, many applications (particularly email and mobile apps) still do not fully support IDNs. There is a need for education and technical development to provide fully localized domain names at the base level for markets around the world.

Richard Ishida presented on the difficulties in dealing with and parsing names around the world. Many developers separate family and given names, but when the entire world is considered, the complexity is enormous and there is a need for more complex ontologies of name structure with appropriate implementations.

Sebastian Hellmann presented on the Linked Open Data 2 (LOD2) stack and its support for the LOD lifecycle. He also discussed the NLP Interchange Format (NIF) and its role in promoting interoperability between NLP tools, resources, and annotations.

Fernando Serván's presentation discussed issues faced in restructuring an existing site to promote multilinguality. As much of the current content on FAO's website is unstructured monolingual content, localizing it and understanding requirements are a major challenge. In addition, in-depth analytics are needed to help prioritize activities to meet user requirements. In the long run standards play a crucial role in promoting reuse of multilingual content.

Paula Shannon closed the session with a presentation on the ways in which international search engine optimization (SEO) needs to go beyond just localizing keywords to include optimization of keywords for specific markets. Companies also need to be aware of ways to build relationships in local markets and avoid problems that arise when content localization is not backed up by actual market presence. These issues are particularly crucial for companies wanting to gain early market share in growing economies like India and China.

The last block in the day consisted of "Open Space" discussion sessions with topics chosen by attendees. These sessions discussed key issues in linked open data, translation quality, internationalized domain names, standards, dealing with names, and strategies for non-market translation.

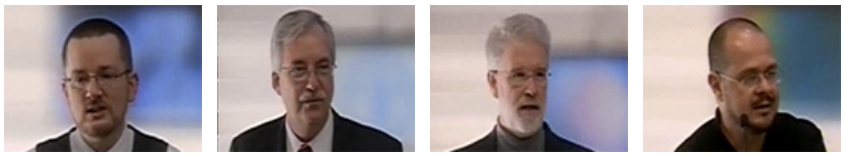
Over the course of the day we heard of many interesting initiatives where standards play a key role. There was a general concern for increasing and maintaining interoperability, and we heard repeated calls for greater public participation in initiatives to move forward the multilingual Web.

The discussions produced a good number of diverse ideas and opinions, and these are summarized at the bottom of this report. There are also links to slides used by the group leaders and video to accompany them, which give more details about the findings of each of the breakout groups.

ITS 2.0 Showcase

As an integral part of the program, the Workshop featured a showcase of implementations of the Internationalization Tag Set 2.0 (ITS 2.0) specification during lunches and breaks. These poster-presentations highlighted the progress made in practical usage of ITS 2.0, which is developed by the MultilingualWeb-LT project, which also administered the Workshop. A list of showcase sessions with links to download the poster files (PDF) is available here.

Welcome session & Keynote talk



Daniel Gustafson, Food and Agriculture Organization of the United Nations (FAO), Deputy Director-General, welcomed the participants to Rome and gave a brief speech ... (encouraging people to come up with ideas that will make it easier to work with the Multilingual Web).

Related links: IRC • Video

This was followed by a brief welcome from **Arle Lommel**, co-chair of the Workshop, who provided an overview of the format and topics of the event.

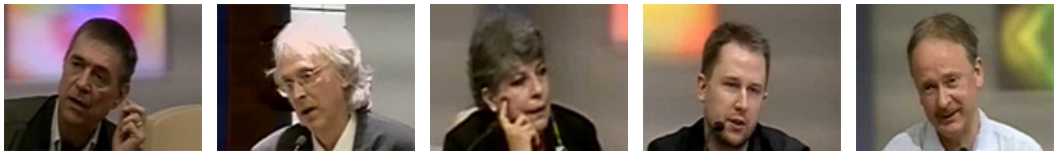
Keynote speakers **Mark Davis** (president of the Unicode Consortium and Internationalization Architect at Google) and **Vladimir Weinstein** (Engineering Manager at Google) spoke about “Innovations in Internationalization at Google.” Their presentation discussed the ways Google, as a company, is working to address the challenge and opportunities presented by multilingual content on the Internet. They discussed how core internationalization features—like Unicode character support, the Common Locale Data Repository (CLDR), and the International Components for Unicode (ICU)—have impacted companies doing business across international boundaries, but that additional work remains. Significant points include:

Related links: Slides • IRC • Video

- Complex issues remain, including dealing with text segmentation, detecting character encodings, working with “entities” and names, phone numbers, addresses, dealing with grammatical complexities of languages, knowing what to translate, working with speech, and dealing with fonts and text input.
- Companies need not only to understand these issues at a technical level, but also how to address complex cultural and security issues that may arrive.
- A large product portfolio means that users experience inconsistent results across various services.
- Google is working on ways to improve Web fonts to eliminate the experience of users encountering “tofu” (the situation where characters are replaced by square blocks because the local operating system lacks fonts to render them).
- There is also a concerted effort to roll out tools that will enable users to create multilingual resources (such as subtitles for YouTube videos) while avoiding the complexity of many proprietary solutions.

Developers session

The developers Session was chaired by **Christian Lieske**, SAP AG



Jan Anders Nelson (Senior Program Manager at Microsoft) began the Developers session with his presentation, “Going Global with Mobile App Development: Enabling the Connected Enterprise.” He discussed Microsoft’s tools for multilingual web and app development (his planned copresenter, Jörg Schütz, was unable to attend the Workshop). The new Multilingual App Toolkit for Windows Phone 8 is built to support standards like XLIFF and to simplify the process of developing apps for a variety of device form factors. Other significant remarks:

Related links: Slides • IRC • Video

- Windows8 currently support 109 languages and Microsoft wants to help developers localize into these languages.
- Support for pseudotranslation (replacement of text with “gibberish” text that mimics the specifics of other languages) enables users of Microsoft’s toolkit to detect problems prior to localization, such as hard-coded, concatenated, or truncated strings, and other visual issues
- The toolkit interface with Microsoft’s translation services to deliver real-time translation and provides XLIFF support.
- The toolkit enables developers to push changes made in one resource to all similar resources (change once, change everywhere)

These tools not only will adjust to various form factors, but also allow cross-platform leverage of app resources between phone and traditional PC.

Gavin Brelstaff (Researcher at CRS4 Sardinia) and **Francesca Chessa** (University of Sassari) presented on "Multilingual Mark-Up of Text-Audio Synchronization at a Word-by-Word Level: How HTML5 May Assist In-Browser Solutions" This presentation described their work to improve audio-text synchronization across languages in the browser using HTML5 capabilities. They had reported on preliminary efforts in this area at the MultilingualWeb Workshop in Pisa in 2011. The solution they developed uses the new Timed Text Markup Language (TTML) specification from W3C, coupled with HTML5's audio capabilities, to provide a simple solution for representing synchronized data. Other significant remarks:

Related links: Slides • IRC • Video

A demonstration of how the solution highlights the correspondence between words in different languages (even where word order is different) and shows correspondence between words, phrases, and abstract ideas.
TTML provides a simple approach representing this data and can be combined with TEI XML to improve interoperability.
In addition, the team has developed a suite of tools for editing time codes in an intuitive, easy-to-use fashion; this tool allows users to make corrections while listening without the need for complex tools.
The goal of the project is to assist with language learning and increase options for learners to use "poetic memory" in understanding languages.

Gábor Hojtsy (Acquia) presented on "Multilingual Challenges from a Tool Developer's Perspective." He spoke about his development experience working with the open-source Drupal content management platform for web communities in Hungary. Because of difficulties he encountered, he has spent a decade working to improve Drupal's multilingual support. Other significant remarks:

Related links: Slides • IRC • Video

Drupal work uses the gettext.po format for localizing the application, which may seem archaic, but which is adequate for their needs. It is used only as a simple transportation format, although other tasks may use other formats.
Drupal.org has over 20,000 modules in a git repository that all take advantage of a unified localization process.
Since some strings span many projects, all strings share a single space in a subdomain, translate.drupal.org, specifically for working on translations.
This site supports "microcontributions," where users can contribute even as little as a single string. There is also a tool for websites to support in-place translation in a website.
Drupal's configuration system is user-editable and modifiable, which poses problems for localization. There is a tool to identify those parts that are editable and translatable so that translate.drupal.org can address them.
There is now a full workflow system that integrates everything and includes support for ITS 2.0 that allows content to be pushed to translators.

Reinhard Schäler, director of the LRC at the University of Limerick, discussed the Rosetta Foundation's first pilot project using SOLAS, a standards-based community space for translation and localization, developed at the Localisation Research Centre at the University of Limerick, in cooperation with the Centre for Next Generation Localisation. This presentation makes a case for demand- and user-driven translation and localization (social localization), and then describes why social localisation requires new technologies—based on open standards and open source—using the experience of the Rosetta Foundation as an example. It demonstrated how SOLAS-Match can be used in this context. Other significant remarks:

Related links: Slides • IRC • Video

Commercial localization can use 15 languages to reach 90% of the world's *customers*, but 6,985 languages are needed to reach 70% of the world's *citizens*, a demographic currently in urgent need of content.
Delivering content to the rest of the world requires new methods that allow people to contribute easily with flexible tools that cannot be achieved with traditional models.
The SOLAS platform, based on open standards, provides a way for people to contribute translations to the "Trommons" (translation commons), where they are accessible and reusable.

The Developers session on the first day ended with a **Q&A** period with questions about the fragmentation of/competition between various localization platforms (versus standardization) and whether it will create problems for users, the emergence of new browser-enabled translation technologies, managing quality in translator communities, the connection of speech technologies and music, the limitations of existing standards, and the need to evolve to address new technologies.

Related links: IRC • Video

Creators session

This session was chaired by **Bryan Schnabel** of Tektronix.







Román Díez González of the Agencia Tributaria (Spanish Tax Agency) and **Pedro L. Díez-Orzas** of Linguaserve spoke about their experience in implementing the Internationalization Tag Set (ITS) 2.0's MT-specific features. It addressed aspects of shifting from HTML 4.01 to HTML5 and strategies for annotating HTML5 content with ITS 2.0 markup in an efficient and pragmatic way, when faced with real-world pressures and requirements. The presentation described how the www.agenciatributaria.es site has been made multilingual using Linguaserve's Real Time Translation System, and the shift HTML5 and experience with ITS2.0 annotation (both automatic and manual). Other significant remarks:

Related links: Slides • IRC • Video

- ITS 2.0 markup improved interoperability across three different MT systems: ATLAS (Linguaserve's Real Time Translation System), Lucy Software MT (Rule-based Machine Translation), and MaTrEx from Dublin City University (Statistical Machine Translation).
- One issue that has to be addressed with using real-time translation is that the content owner loses control over the translation process. ITS 2.0 markup, however, enables the owner to regain some degree of control through content markup.
- ITS 2.0 requires HTML5. The conversion from HTML 4.01 to HTML5 can be complex, but Linguaserve was able to develop a "shallow" conversion that addressed some of the more problematic aspects automatically.
- Often legacy content has custom ways to address certain needs: mapping to ITS2.0, where relevant, provides a good mechanism to standardize these solutions so that content owners can take advantage of standards-based processes.
- Linguaserve developed components to allow authors to apply manual markup to new content.

Hans-Ulrich von Freyberg, CEO of Cocomore spoke on how standard can help drive business applications forward. Although still in development, the ITS 2.0 standard (developed by the W3C's MultilingualWeb-LT project) is already proving that it can fulfill this promise. Freyburg's presentation showed how the German Industrial Machine Builders' Association (VDMA) and Cocomore, as its service provider, benefit from the development of the ITS 2.0 standard. It demonstrated how the systems created during the standardisation effort support developing client relationships and business opportunities. As a further aspect, it is shown how the results of the standard development process have impacted VDMA's ability to conserve valuable resources. Other significant remarks:

Related links: Slides • IRC • Video

- Managing translation currently can take too much time for many important materials, especially in competitive markets.
- Cocomore has developed a set of ITS 2.0 rules specifically for Drupal, which required extension of the Drupal Translation Management Tool (TMGMT) to support Linguaserve translation and content enrichment through JSI Enrycher. This extension allows authors to: identify content that should not be translated, specify for which locales content should be translated, identify text domain (subject field), add localization notes, engage in text analysis (e.g., named entity extraction), and specify language settings.
- After text has been marked up by authors, it is exported to XML using ITS markup to preserve the metadata, which is then passed on to language (translation) services.
- Early results show that using ITS 2.0 results in a 30% reduction in delivery times for translated texts, e.g., a four-hour project might be delivered in 2:45. Most of the reduction comes in reduced project management, annotation, and round tripping tasks. Moving to MT could deliver even more savings.

Brian Teeman, co-founder of Joomla, spoke on how Joomla!, used by over 2.8% of the web and by over 3000 government web sites, is working to make web sites truly multilingual, rather than relying on automated translation tools. The latest release of Joomla makes building multilingual sites considerably easier,

Related links: Slides • IRC • Video

while also making them more accessible across different user agents and form factors (e.g., desktop or laptop computers, mobile phones, or tablets). This presentation showcased how these new developments can be used in Joomla to greatly reduce the burden of building and releasing multilingual sites. It focused on the ways that Joomla has worked to simplify making sites independent of form factor and has extended the same philosophy to multilinguality. Other significant remarks:

- Joomla is working to build integration with community translation services to help content owners access resources for translation.
- Joomla users can install Josetta to obtain a Joomla-based translation management system experience.

Vivien Petras, Humboldt Universität zu Berlin, addressed the semantic and multilingual interoperability challenges that arise when building large-scale cultural heritage information systems. She presented Europeana—the European digital library, archive, and museum—as a use case to discuss concrete processes and issues that arise when developing an aggregated solution to access very heterogeneous collections of cultural heritage material. The presentations highlighted issues that arise in targeting a "European" audience. It also covered how sparse metadata in different formats can be aggregated and enriched in order to provide a satisfactory multilingual user experience. Other significant remarks:

Related links: Slides • IRC • Video

- Although most Europeana objects are language independent, the metadata describing them is monolingual, presenting problems for identifying relevant resources, especially when queries are run in a less common language.
- 69% of site visitors use a native-language browser and the most common entry point to Europeana is a native-language Google search (91% of visitors). As a result the site must support local language requirements.
- Using off-the-shelf MT systems like Microsoft translation creates problems because such systems are not adapted to specific domains, but most Europeana content is highly specific to a domain.
- Automatic content enrichment can deliver incorrect results and needs better input. Ambiguity is a major source of problems, especially when language identification data is inadequate and words have different meanings in different languages. Linked open data that refers to authoritative contextual vocabularies is providing one way to address these issues effectively.

Christian Lieske and **Inna Nickel** of SAP reported on "Tool-Supported Linguistic Quality in Web-Related Multilanguage Scenarios." This report was based on collaboration with Daniel Naber from LanguageTool. Two main questions were addressed: (1) why is tool support needed?, and (2) what tool support does already exist? A short introduction into Natural Language Processing (NLP) was provided. NLP facilitates linguistic quality of textual content—correct spelling, terminology, grammar, and style—and is thus of utmost importance for various content-related processes in the realm of linguistic quality. Humans, as well as search engines and Machine Translation systems, for example, benefit from high-quality content. After providing a brief overview of LanguageTool, its use in scenarios involving writing guidelines for Russian and checking compliance with "Leichte Sprache" (easy language) in German-speaking countries (developed with Annika Nietzio) were discussed. Other significant remarks:

Related links: Slides • IRC • Video

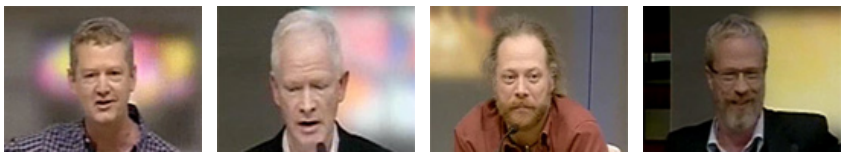
- Content-related quality has major dimensions such as "technical" and "linguistic" and enterprises and public service may have different views on linguistic quality. Public services in particular are seeing an increasing demand to produce content that allows more inclusion (e.g., via Easy-to-Read content)
- Very often, production of content for the Web involves many actors/handovers, has to be fast, and has to cope with huge volumes, making it necessary to quickly identify problems in the text.
- LanguageTool's feature set makes it very attractive for this sort of work: It has support for approximately 30 languages, performs language-independent as well as bilingual checks, has support for the draft W3C Internationalization Tag Set 2.0 (for some languages), is integrated with LibreOffice/OpenOffice and Okapi-based applications, it can be run as a server that can be reached via HTTP(S), and is available as a plug-in for Firefox browser

The **Q/A** part of the Creators session addressed questions about support for terminology management in CMS, requests for support for ITS 2.0 in Joomla, alternative way to achieve integration of LanguageTool, extension of CMS to support localization, and clarification on where savings from technology can be achieved.

Related links: IRC • Video

Localizers session

This session was chaired by **Jan Nelson** of Microsoft.



Bryan Schnabel, Content Management Architect at Tektronix and Chair of the XLIFF Technical Committee, discussed how an emerging standards-based, best-practice approach to working with CMS that can deliver significant time and cost savings. This presentation demonstrated how the use of a CMS, combined with open standards, moved Schnabel's team one step closer to a turn-key multilingual web workflow. It showed how using a component CMS (Trisoft, with DITA and XLIFF), and web CMS (Drupal, and the Drupal XLIFF module) improves quality, reduces cost, and reduces time-to-market. It also addressed the architecture, the hurdles, and the benefits experienced. Other significant remarks:

Related links: Slides • IRC • Video

- The existing Drupal solution is easy to use, but it doesn't allow the translator to access translation memory (TM) and the translator can actually break things. The need, therefore, was to leverage XLIFF to improve Drupal.
- With the XLIFF scenario, Drupal can save text nodes to XLIFF, which can be sent on to the LSP for translation. This simplifies using TM and other translation resources accessible to the LSP. Not all text (e.g., UI strings) can be saved directly as XLIFF, so PO files are needed in some cases as an intermediary. This approach is better for the translator than the native Drupal methods.
- DITA has many advantages for content management, but without XLIFF it requires sending many, many files. By using XLIFF integrated in Trisoft via the XLIFF dita open toolkit plugin, many DITA topics can be converted into one XLIFF file, simplifying management tasks for content owners and translators.

Sinclair Morgan, SDL, discussed the challenges faced by today's public organizations and institutions in creating fluid, highly customized, and on-demand information across multiple channels and geographies. He discussed real-life examples to demonstrate how integrated machine translation and post-editing is in use to considerably increase the amount of multilingual information that is being published on the web, without compromising on quality. Other significant remarks:

Related links: Slides • IRC • Video

- The effectiveness of MT + postediting versus human translation varies considerably, but there is typically a reduction in cost and time to delivery of 30% compared to conventional translation with translation memory
- For postediting to be effective, the MT system needs to be adapted to specific texts, domains, and tasks in an ergonomic and intuitive technology platform.
- The same quality checks used for human translation need to be applied to MT.

Charles McCathie Nevile discussed "Localization in a Big Company: The Hard Bits." He discussed the large technology portfolio used by Yandex to deliver its content, including machine-assisted translation systems, localisation systems, content management systems, and front-end development modularisation. Despite all of this technology, things can still go wrong and considerable work remains to deliver better multilingual content. Language knowledge isn't all either, as companies need to be aware of many locale-specific cultural requirements. Other significant remarks:

Related links: Slides • IRC • Video

- Yandex started to use Russian morphology tools to improve search results over what was then possible.
- Some of Yandex's services are specific to individual markets, e.g., Russia and Kazakhstan. However, they have found that extensive personalization is not appreciated by customers.
- Companies need to be aware of their internal biases and need to work to overcome them. Expectations based on language can hinder companies in understanding and addressing issue that affect other markets.

Hans Uszkoreit, DFKI, spoke on "Quality Translation: Addressing the Next Barrier to Multilingual Communication on the Internet." In his presentation he discussed how language is still one of the most pervasive barriers to interpersonal communication, cross-border commerce, and full participation in European democracy, even as Europe has seen political integration. Despite recent progress in machine translation, it is clear that the quality of today's Internet-based translation services is neither good enough for many tasks nor complete in terms of language coverage. Speakers of smaller languages thus find themselves largely excluded from vital discussions of European identity and policy. His talk argued in favor of a concerted push in Europe for quality translation technology and addressed concrete preparatory actions performed by the EC-funded QT LaunchPad Project. Other significant remarks:

Related links: Slides • IRC • Video

- Demands for translation will require MT, but current MT research has focused primarily on in-bound translation requirements where quality is not critical. As a result MT solutions are lacking for many language pairs and for many domains.
- For outbound translation, we need better ways to identify what is good enough, almost good enough, and not usable.

Unfortunately most MT improvement is happening in the “not usable” range.

The QTLaunchPad project is looking at ways to move “almost good enough” into “good enough.”

The related META-NET project produced a strategic research agenda on three research topics: the translingual cloud, social intelligence, and socially aware interactive assistants. All of these rely on improvements in translation quality.

The QTLaunchPad project is working on shared quality metrics for human and machine translation.

During the **Q&A** session questions were raised about using proxy servers to simplify language deployment, costs of tuning MT, MT technology types, the diversity of translation quality requirements, standardization of translation packages, specialization of translation services in the cloud, determining what quality matters, language dependence of knowledge, and importance of quality.

Related links: IRC • Video

Machines session

This session was chaired by **Feiyu Xu** of DFKI.





José Emilio Labra Gayo, University of Oviedo, talked about "Multilingual Linked Open Data Patterns". He presented a catalog of patterns and best practices to publish multilingual linked data. Each pattern contains a description, a context, an example and a short discussion of its usage. Other significant remarks:

Related links: Slides • IRC • Video

Patterns are classified by activity around multilingual linked open data, like naming URIs, dereferencing or linking.

Long descriptions can be divided to provide more readable information in several languages and to leverage existing Internationalization techniques like markup for bidirectionality, Ruby or translation information.

So-called “soft inter-language links” can help to link between multilingual datasets, without implying strong semantic equivalence of concepts across languages.

Reuse of existing, wide spread vocabularies (FOAF, Dublin Core…) helps to control vocabulary evolution, but it requires careful localization of the vocabularies.

Asunción Gómez-Pérez Universidad Politécnica de Madrid, spoke about the "Multilingualism in Linked Data". The generation of multilingual linked data involves several steps: selection of data sources, modelling of the underlying information domain, transformation into RDF, linking between the new RDF data set and existing ones, and publication: of the model, the RDF resources, and links. The published data then can be exploited in applications. Multilingualism is present in all parts of this linked data life cycle, from specification through to maintenance and use. Other significant remarks:

Related links: Slides • IRC • Video

Links across languages can be discovered at runtime or offline; they can be stored in separate repositories for semantic applications like cross-language question answering.

While exploiting the multilingual data, users should be able to pose questions in their own language to be processed against the Web of linked data.

To realize the multilingual Web of data, several services are needed: for multilingual linked data generation, for translation / localization, cross-lingual linkage, and for cross-lingual access.

Peter Schmitz, Publications Office EU, gave a talk about "Public Linked Open Data - the Publications Office´s Contribution to the Semantic Web". The Publications Office of the European Union contributes to the Semantic Web via various efforts: the CELLAR repository exposes metadata about official EU information. The open data portal provides access of huge data sets as linked open data. The Publication Office also participates in standardization, contributing to the European Legislation Identifier (ELI). Finally, the Publications Office provides multilingual controlled vocabularies for re-use. Other significant remarks:

Related links: Slides • IRC

Multilingualism is core business of the Publications Office, daily publishing in up to 23 languages including various multilingual web sites.

An RDF based interface is provided to upload data to the open data portal. Already several data providers submit RDF directly to the portal.

The Interinstitutional Metadata Registry provides translations of controlled vocabularies that are discussed between many European institutions. This leads to high translation quality and broad consensus across Europe member states.

Gordon Dunsire, Independent, gave a talk about "Multilingual Issues in the representation of International Bibliographic Standards for the Semantic Web". The International Federation of Library Associations and Institutions (IFLA) maintains standards for the library and bibliographic environment. Current initiatives to apply the IFLA language policy to linked data include guidelines on translations of namespaces, the Multilingual Dictionary of Cataloguing, and current multilingual element sets and value vocabularies. These include the Functional Requirements family of models and International Standard Bibliographic Description. Other significant remarks:

Related links: Slides • IRC • Video

IFLA has seven official languages, but translations are undertaken for many more languages.

The open metadata registry is used to store element sets and value vocabularies as RDF.

When translating the opaque URIs in the registry, various issues come up: what to translate first, what style to use, language inflection, disambiguation of labels etc.

IFLA is working on guidelines to describe the issues and related best practices.

Thierry Declerck, DFKI, gave a talk about "Language Technology Tools for Supporting the Multilingual Web". He presented joint work by himself and Max Silberztein, the main developer of the NooJ linguistic development environment. Around this tool set a strong community has emerged, with language resources for 22 languages available, and more in development. NooJ provides robust finite state tools for generating annotated multilingual resources for the web, including the ability to process HTML/AML-annotated documents and to transform those annotations for specific purposes such standardised markup for supporting multilingual applications. Other significant remarks:

Related links: Slides • IRC • Video

Multilingual Semantic Resources are already available on the Web, but some of these first have to be mapped to RDF to be used in linked data scenarios.

An application of NooJ using multilingual resources takes ontology or taxonomy elements together with preferred and alternative labels and automatically generates a grammar.

The grammar can be used for automatic annotation of text using NooJ or other annotation frameworks.

The discussion during the **Q&A** session centered around the topic of adoption and the relation to other communities and their (multi- or monolingual) resources: how to make the ordinary (linked open data) Web developer aware of issues and benefits from multilingual linked open data? The W3C Ontolex community group is already contributing in this direction. Further steps could include approaching the Schema.org community about providing multilingual labels + documentation. But also a strong effort of writing best practices + creating roadmaps for long-term, wide spread adoption is needed.

Related links: IRC

One issue with multilingual linked open data is that even the best practices discussed during this session are not stable yet. So when linked open data is published, often language related issues are only an afterthought. Here, services that ease the task of adding translations to linked data could help. In such scenarios the role of content author becomes critical, and the ability to identify items as as result of an automated annotation process. As a pre-requisite for learning multilingual ontologies, the resource Freebase can serve as an interlingua.

Users session

This session was chaired by **Richard Ishida** of W3C.

Pat Kane, Verisign, started the Users session with a talk entitled "Internationalized Domain Names: Challenges and Opportunities". Today International Domain Names (IDNs) are getting more and more attention, but are far from being ubiquitous and trusted. To change the picture, registries, developers, content creators, policy and standard making bodies all need to work together on internationalization of the identifiers on the Internet. Pat Kane presented challenges Verisign has found to make IDNs a ubiquitous and trusted part of the multilingual Web. Other significant remarks:

Related links: Slides • IRC • Video

- Currently about 1.3 millions are registered in .com and .net, with a focus on IDNs with Chinese-Korean-Japanese (CJK) ideographs.
- To foster wide spread adoption of IDNs, support in applications is crucial: browsers, email addresses & clients, esp. mobile applications etc.
- Users in many communities want IDN.IDN (that is, IDNs both in the top level domain and other parts of the domain). However, surveys show that businesses won't buy IDN.IDN, because of lack of ubiquity, trust and awareness.
- Depending on the language community - e.g. China versus India versus Korea -, the interest in IDN.IDN or ASCII only domain names is quite different.

Richard Ishida, W3C, talked about "What´s in a Name?". The presentation showed various issues that arise when dealing with names in forms and ontologies. It is difficult to generalize the issues across languages and to find a general solution; language- and culture-specific knowledge is needed to represent names adequately. The talk demonstrated the problems using examples from a wide range of languages. Other significant remarks:

Related links: Slides • IRC • Video

- People defining forms or ontologies need authoritative information about what fields they should be using.
- They also need authoritative guidance about how to approach the problems of handling personal names.
- This guidance has been needed for some time, and it would be good to see some work on those issues.
- There are similar issues and requirements for dealing with things like addresses.

Sebastian Hellmann, Leipzig University, talked about how "The LOD2 Stack and the NLP2RDF Project". The LOD2 stack provides tools for all parts of the linked data life cycle: from extraction of linked data, through exploration and query, to knowledge enrichment as a basis for search and further exploration. An important part of the stack are components to process natural language content. Here the NLP Interchange Format (NIF) comes into play. NIF fosters interoperability between Natural Language Processing (NLP) tools, language resources and annotations. Other significant remarks:

Related links: Slides • IRC • Video

- One key achievement of the LOD2 project is that the tools are provided on the LOD2 stack homepage in a robust, ready to use manner.
- The wikipedia article about knowledge extraction developed with contributions from LOD2 provides a lot of information about both extraction approaches and tools.
- The version 2.0 of NIF is to be completed within 2013.

Fernando Serván, Senior Programme Officer at the Food and Agriculture Organization of the United Nations, talked about how "Reorganizing Information in a Multilingual Website: Issues and Challenges". FAO is an international organization with more than 190 member countries and 6 official languages. The presentation discussed issues and challenges faced when reorganizing the content of the FAO multilingual website. Sample challenges are working with existing, unstructured multilingual content, or the development of ontologies to navigate and map new content. Other significant remarks:

Related links: Slides • IRC • Video

- Millions of users visit FAO's web presence and ask for content in their language.
- Analytics are important for decision-making and prioritization.
- Standards play an important role for assuring long-term use of multilingual content.

Paula Shannon, Lionbridge, talked about how "The Globalization Penalty". Search engine optimization (SEO) is key for a successful web presence. The presentation discussed the role of the translation process for SEO. That process needs to go beyond simple keyword localization to achieve global web site traffic. Several case studies showed the importance of this development. The main aspects of SEO for global markets are global search performance, international keyword optimization, mapping of content and keywords, and the overall management of internationalized SEO. Other significant remarks:

Related links: Slides • IRC • Video

- Branding in the digital age is very different: younger consumers will not care of how products are advertised, they will search for

them and then decide.

In this situation, companies need to be global but also win on the local search field.

Three business issues in digital branding are: not enough traffic, page visits not leading to customer relations, and management of a multilingual website.

Mastering these issues is crucial for success in growing online markets like India or China.

Related links: IRC • Video

In the **Q&A** session the growing importance of the user perspective was discussed. Companies have to stop the traditional "push" approach of advertising their products. Rather, they need to change to a "pull" approach and carefully listen to potential customers on the Web. Tools to build the bridges between companies and customers involve technologies like (social media) analytics. But their applicability depends on the region in question and cannot be easily generalized world-wide.

An example of the region specific applicability are IDNs. One has to study carefully user behaviour and needs in online markets to find the right approach for IDN adoption. One challenge for localization is that in most companies technology is mostly about "doing more with less " and cost savings. Nevertheless some enterprise companies have started to focus also in localization in users needs and take these into account e.g. when developing terminologies.

Discussion sessions

Related links: IRC

This session was chaired by **Des Oates** of Adobe.

Workshop participants were asked to suggest topics for discussion on small pieces of paper that were then stuck on a set of whiteboards. Des then lead the group in grouping the ideas and selecting a number of topics for breakout sessions. People voted for the discussion they wanted to participate in, and a group chair was chosen to facilitate. The participants then separated into breakout areas for the discussion, and near the end of the workshop met together again in plenary to discuss the findings of each group. Participants were able to move between breakout groups.

In Rome we split into the following groups:

1. Best Practices for Multilingual Linked Open Data (BP-MLOD)
2. Translation Quality
3. Internationalized Domain Names (IDNs)
4. Standards
5. Names
6. Crowdsourcings/non-market strategies

The summary of the group discussions was followed by a general discussion and wrap up. Key findings of the groups are provided below, some of which have been contributed by the breakout group chairs.

Best Practices for Multilingual Linked Open Data (BP-MLOD)

Related links: Slides • IRC • Group report

This session encompassed a number of presentations and was summarized by José Emilio Labra Gayo, University of Oviedo. In various areas more guidance is needed, e.g. about the question whether to use URIs or IRIs (Internationalized Resource Identifiers), or to use opaque or descriptive URIs. One best practice that had brought support was to recommend language tags in linked open data.

Linkage between languages within linked data can be realized with the `owl:sameAs` construct. However it implies very strong semantic

relations. As an alternative, the notion of "soft links" e.g. realized via SKOS came up.

The general feeling about this session was that there is a critical mass of people to continue work on the topic. As one major outcome of the Rome workshop, a forum for the continuation was created: the Best Practices for Multilingual Linked Open Data W3C community group. As of writing, there are already 46 participants. This is an *open forum*: joining the group requires only a W3C account.

Gordon Dunsire: “Multilingual bibliographic standards in RDF: the IFLA experience” (video)

Ivan Herman: “Towards Multilingual Data on the Web?”(video)

Jose E. Labra: “Patterns for Multilingual LOD: an overview”(video)

Dave Lewis: “XLIFF workflow and Multilingual Provenance in Linked Data”(video)

Charles McCathie Nevile (video)

Roberto Navigli: "BabelNet: a multilingual encyclopedic dictionary as LOD" (video)

Haofen Wang: “The state of the art of Chinese LOD development” (video)

Daniel Vila: “Naming and Labeling Ontologies in the Multilingual Web” (video)

Translation Quality

Related links: Slides • IRC • Video

This session was summarized by Arle Lommel, DFKI. For identifying quality issues, it is important not to focus on the production methods used in translation (human, machine translation, ...) but rather take the perspective of the end user / consumer of the translation. In addition, the quality of the source needs to be assessed. To put it differently, the definition of quality metrics depends on which step is analyzed: content production, translation, revision, etc.

Many metrics to evaluate translation quality are available, but applying the metrics often does not lead to reproducible results; the metrics are rather academic and not usable in real life production scenarios. When applying metrics in translation workflows, a feedback loop is needed to avoid repetition of errors.

To move the topic of translation quality forward, many parties need to be involved: translators, post editors, and various business worlds. The MultilingualWeb community can help to create a forum for discussing the topic and to help avoiding the "reinvent the wheel" syndrome, e.g. by bringing quality metrics work undertaken within the QTLaunchPad project and TAUS together. However, so far no concrete umbrella effort was started as the result of the Rome workshop.

Internationalized Domain Names, (IDN)

Related links: IRC • Video

This session was summarized by Pat Kane, from Verisign. Pat made clear that the adoption of IDN is still an ecosystem problem. Users will embrace IDN only if they can rely on them in mail clients, web browsers & servers, all other parts of the internet and Web infrastructure, but also outside the Web.

To move the topic forward, key players in above (tool) areas need to be brought together. W3C might be the place to achieve this community building; however, it is important to engage a wide range of groups, including both technical sub areas and political decision makers. Pat said that Verisign will start participating in several W3C working groups to explore the potential of W3C to become a platform for fostering IDNs world wide.

Standards

Related links: IRC • Video

This session was summarized by David Filip, University of Limerick. David first mentioned various usage scenarios for standardization in a localization workflow. Standards need to ease the creation of a general content management localization roundtrip, the application of terminology management, and general tool integration. The good news is that most of these and other use cases are covered by XLIFF 2.0 and ITS 2.0. For example, new ITS 2.0 data categories like MT Confidence, Provenance or Text Analysis ease the creation of machine translation workflows or terminology evaluation.

A next step might be not to put more effort in format standardization, but focus on defining interfaces e.g. in the field of terminology messaging. These could help to facilitate terminology exchange. In the discussion about the session also a warning came up: in localization there is a proliferation of standards, so one should carefully consider whether new standardization is really needed.

Names

Related links: Slides • IRC • Video

This session was summarized by Juan Pane, University of Trento. The group had identified three tasks related to names. The first is recognition of names. This encompasses sub tasks like named entity recognition or machine translation, in order to verify whether two items are

the same across languages. Only with such a verification, applications like multilingual business intelligence can be created.

The second task is display of names. Display issues need to be resolved for sorting, contextual usage or text-to-speech generation. The third task related to names is capturing. This might involve transliteration, speech-to-text and handling of input from forms.

The focus in this session was on personal names. Other types of names e.g. for organizations or events were not discussed in detail. One concrete next step around the topic could be to come up with a standardized ontology. The purpose would be to identify all possible formats of names.

Crowdsourcing/non-market strategies

Related links: IRC • Video

This session was summarized by Reinhard Schäler, director LRC, University of Limerick. The main discussion was about how to get people involved in volunteer translations. In normal life, translation is a business for money. What incentives are needed to gather enthusastic, voluntary translators? And even if there is a critical mass of voluntary translators, challenges remain to be resolved, like the translation tooling setup. Volunteers can only be involved in translation processes if the translation environments are easy to use.

The participants of the session looked into concrete projects where volunteer translators could help, like the FAO Web site development, and into tooling that can support them, like SOLAS. There is a desire to bring user needs and tooling together in real projects; however, no concrete action was planned during the session.

Open Space Q&A

Related links: IRC • Video

The workshop was closed with a general discussion, covering all breakout sessions. Many resources needed for multilingual processing are closely related: linked open data, terminologies, lexicons etc. There seems to be a need to bring efforts togehter, detect gaps and overlaps. This will help to avoid the "reinventing the wheel" syndrome.

To achieve this kind of broad consensus building, new fora like the W3C Best Practices for Multilingual Linked Open Data Community Group should make an effort from the start to engage both cooperate users from industry and research. GALA can play a key role for involving the localization industry.

During the week after the Rome workshop, several players from the MultilngualWeb community participated in the GALA conference and spread the word about needs of the multilingual Web. Sending a clear message also as part of such events is crucial, so that gaps and needs will be recognized by the right group of people. Only in this way we can expect tangible actions to foster the adoption of the MultilingualWeb.

Authors: Arle Lommel, Nieves Sande, Felix Sasaki. Contributors: Scribes for the workshop sessions.

Photos in the collage at the top and various other photos, courtesy FAO, Arle Lommel, Richard Ishida, Jan Nelson, and Wikimedia Commons. Video recording by Food and Agriculture Organisation, and hosting of the video content by VideoLectures.