2nd workshop
Content on the Multilingual Web
4-5 April 2011

MultilingualWeb

Area della Ricerca CNR
Via G. Moruzzi, 1
56124 Pisa, Italy

4 April 2011

WELCOME

9:00

Domenico Laforenza – CNR-IIT
The Italian approach to Internationalized Domain Names (IDNs)

Basically this is a system through which you can "write" on the Internet, for example in Danish or Chinese, using accented letters or non-Latin characters. Until recently, the choice of domain names was limited by the twenty-six Latin characters used in English (in addition to the ten digits and the hyphen "-"). IDN, introduced by ICANN (Internet Corporation for Assigned Names and Numbers) represents a breakthrough, for hundreds of millions of Internet users in the world that until now were forced to use an alphabet that was not their own. With regard to Italy, the impact of accents will certainly be less marked, but it will give everyone the opportunity to register domains which completely match the name of the person, company or brand name chosen.

9:20

Oreste Signore – CNR / W3C Italy
Is the Web really a "Web for All"?

This talk will be a brief recall of the basic issues of the Web: multicultural, multilingual, for all. A look to the relevant W3C activities to pursue the ultimate goal of One Web.

9:35

Kimmo Rossi - European Commission – DG INFSO E1
Brief welcome address

9:50

Ralf Steinberger – European Commission - Joint Research Centre (JRC)
Complementarity of information found in media reports across different countries and languages

There is ample evidence that information published in the media in different countries is largely complementary and that only the biggest stories are being discussed internationally. This applies to facts (e.g. on disease outbreaks or violent events) and to opinions (e.g. the same subject may be discussed with very different emotions across countries), but there is also a more subtle bias of the media: National media prefer to talk about local issues and about the actions of their politicians, giving their readers an inflated impression of the importance of their own country. Monitoring the media from many countries and aggregating the information found there would allow readers a less biased and more equilibrated view, but how to achieve this aggregation? The speaker will give evidence of such information complementarity from the Europe Media Monitor family of applications (accessible at http://emm.newsbrief.eu/overview.html) and show first steps towards the aggregation of information from highly multilingual news collections.

KEYNOTE

Q&A

DEVELOPERS

11:00

Steven Pemberton – CWI/W3C
Multilingual forms and applications

This talk will describe the use of XForms to simplify the administration of multi-lingual forms and applications. A number of approaches are possible, using generic features of XForms, that allow there to be one form, with all the text centralised, separate from the application itself. This can be compared to how style sheets allow styling to be centralised away from a page, and allow one page to have several stylings; the XForms techniques can provide a sort of Language-Sheet facility.

Marcos Caceres – Opera Software
Lessons from standardizing i18n aspects of packaged web applications

The W3C's Widget specifications have seen a great deal of support and uptake within industry. Widget-based products are now numerous in the market and play a central role in delivering packaged web applications to consumers. Despite this, the W3C's Widget specifications, and its proponents, have faced significant challenges in both specifying and achieving adoption of i18n capabilities. This talk would describe how the W3C's Web Apps and i18n Working Group collaborated to create an i18n model, the challenges we faced in the market and within the W3C Consortium, and how some of those challenges were overcome. This talk would propose some rethink of best practices and relay some hard lessons learned from the trenches.

Richard Ishida – W3C
HTML5 proposed markup changes related to internationalization

HTML5 is proposing changes to the markup used for internationalization of web pages. They include character encoding declarations, language declarations, ruby, and the new elements and attributes for bidi support. HTML5 is still very much work in progress, and these topics are still under discussion. The talk aims to spread awareness of proposed changes so that people can participate in the discussion.

Gunnar Bittersmann – VZ Netzwerke
Internationalization (or the lack of it) in current browsers

I'll address two common i18n problems that users of current mainstream browsers face. Users should get content from multilingual Web sites automatically in a language they understand, hence they need a way to tell their preferences. Some browsers give users this option, but others don't. I'll demonstrate live if and how languages can be set in various browsers and discuss the usability issue that browser vendors

have to deal with: the trade-off between functionality and a simple user interface. Users should also be able to enter email addresses with international domain names into forms. That might not be possible in modern browsers that already support HTML5's new email input type. I'll show how to validate email addresses not being too restrictive and eventually raise the question: Does the HTML5 specification have to be changed to reflect the users' needs?

Jochen Leidner – Thomson Reuters
What's Next in Multilinguality, Web News & Social Media Standardization?

The Web is no longer just a protocol (HTTP) and a mark-up language (XHTML); rather, it has become an ecosystem of different content mark-up standards, conventions, proprietary technologies, and multimedia (audio, video, 3D). The static Web page is no longer the sole inhabitant of that ecosystems: Web applications (from CGI to AJAX), Web services, and social media hubs with huge transaction volumes that exhibit some properties of ITsystems and social fabric. In this talk, I would like to discuss some of the challenges that this diversity implies for the technology and stack, to assess the standardization situation, and to speculate what the future may (and perhaps should?) bring.

Q&A

14:00

CREATORS

Dag Schmidtke – Microsoft European Development Centre
Office.com 2010: Re-engineering for Global reach and local touch

Office.com is one of the largest multilingual content driven web-sites in the world. With more than 1 billion visits per year, it reaches 40 languages. For the Office 2010 release, authoring and publishing for Office.com was changed to make use of Microsoft Word and SharePoint. A large migration effort was undertaken to move 5 million+ assets for 40 markets to new file formats and management systems. In this talk we will present lessons learnt for designing and managing multilingual web-sites from this major re-engineering exercise.

Jirka Kosek – University of Economics, Prague
Using ITS in the common content formats

Internationalization Tag Set (ITS) is set of generic elements and attributes which can be used in any XML content format to support easier internationalization and localization of documents. In this talk examples and advantages of using ITS in formats like XHTML, DITA and DocBook will be shown. Also problems of integration with HTML5 will be briefly discussed.

Serena Pastore – INAF
Obstacles for following i18n best practices when developing content at INAF

This talk addresses the following: a. what could be the best way to produce multilingual web content to comunicate astrophysical science and projects; b. how we could educate and persuade our creators to follow internationalization while using their preferred web authoring tools or web content management systems.

Manuel Tomas Carrasco Benitez – European Commission, Charles McCathieNevile – Opera Software
Standards for Multilingual Web Sites

Additional standards are required to facilitate the use and construction of multilingual web sites. The user interface standards should be a best practices guide combining existing mechanisms such as transparent content negotiation (TCN) and new techniques such as a language button in the browser. Servers should expect the same API to the content, though eventually one should address the whole cycle of Authorship, Translation and Publishing Chain (ATP-chain).

Sophie Hurst – SDL
Local is Global: Effective Multilingual Web Strategies

Web on-the-go is now an everyday reality. It touches all of our lives from the moment we wake, to our commute, from work to an evening out on the town. This reality presents both an opportunity and an incredible challenge as Web content managers attempt to optimize customer engagement. Because visitors do not see themselves as part of a global audience but as individuals, we will examine the WCM software requirements that enable organizations to maintain central control, while providing their audiences with locally relevant and translated content. From a Global Brand Management perspective, we will examine how organizations can manage, and build and sustain a global brand identity by reusing brand assets across all channels (multiple, multilingual websites, email and mobile websites). We will also take a fresh look at automated personalization and profiling, and how Web content can be targeted for specific language requirements as well as the local interests of local audiences.

Q&A

LOCALISERS

16:30

Christian Lieske – SAP, Felix Sasaki – DFKI, Yves Savourel – Enlaso
The Bricks to Build Tomorrow's Translation Technologies and Processes

Although support for standards such as XLIFF and TMX has increased interoperability among tools, today's translation-related processes are facing challenges beyond the ability to import and export files. They require standards that are granular and more flexible. Using concrete examples of the ways that various tools can interoperate beyond the exchange of files, this session walks through some of the issues encountered and outlines the use of a new approach to standardization in which modular standards that, similar to Lego® blocks, could serve as core components for tomorrow's agile, interoperable, and innovative translation technologies.

David Filip – LRC / CNGL / University of Limerick
Multilingual transformations on the web via XLIFF current and via XLIFF next

I will argue that content metadata must survive language transformations to be of use in multilingual web. In order to achieve that goal, content creation and content langauge transformation related meta-data must be congruent, i.e. designed upfront with the transformation processes in mind. To make the point for XLIFF as the principal vehicle for critical metadata throughout multilingual transformations, it will be necessary to give a high level overview of XLIFF structure and functions, both in the current version and the next generation standard that is currently a major and exciting work in progress in the OASIS XLIFF TC.

Sven C. Andrä – Andrä AG
Interoperability Now! A pragmatic approach to interoperability in language technology

Existing language technology standards give the false impression of interoperability between tool. There's a gap to bridge that is mostly about mindsets, technology and mutual consent on the interpretation of standards. A couple of players agreed to search for this mutual consent based on existing standards to bridge this gap. The talk will give some background on the issues with the use of existing standards and how Interoperability Now! is approaching this.

Eliott Nedas – XTM International
Flexibility and robustness: The cloud, standards, web services and the hybrid future of translation technology

First 5 minutes: Introducing the current state of affairs, describing leading innovations. Also lamenting the demise of LISA. Second 5 minutes: Describing the possible future and who will be the winners, who will be the losers. Last 5 minutes: What we can do to get standards moving internally in medium, large, organisations.

Pål Nes – Opera Software ASA
Challenges in Crowd-sourcing

Opera Software has a large community, with members from all over the world. The talk will present various obstacles encountered and lessons learned from using a community of external volunteer resources for localization in a closed-source environment. Included topics will be training and organization of volunteers and managing terminology and branding, as well as other issues that come with the territory.

Manuel Herranz – PangeaMT – Pangeanic
Open Standards in Machine Translation

The web is an open space and the standards by which it is "governed" must be open. However, one barrier clearly remains to make the web even more transnational and truly global. This has been called "the language barrier". Language Service Providers translation business model is clearly antiquated and it is increasingly being questioned when we face real translation needs by web users. Here, immediacy is paramount. This talk is about open standards in machine translation technologies and workflows, supporting a truly multilingual web.

David Grunwald – GTS Translation
Website translation using post-edited machine translation and crowdsourcing

GTS has developed a plugin for websites developed using the open-source Wordpress CMS. It is the only solution that supports post-editing MT and allows content publishers to create their own translation community. This talk will present our system and describe some of the challenges in translation of dynamic web content and the potential rewards that our concept holds.

Q&A

20:00
Evening Reception
Chiostro di San Francesco

To further promote networking among attendees, there will be a reception at 8pm in the Capitolium Hall of the Chiostro di San Francesco, a wonderful ancient cloister, next to the church of St. Francesco.

Entry is free to workshop participants. The Capitolium Hall has frescoes by Niccolò di Pietro Gerini with Histories of the life of Christ (1392). The rectangular cloister is from the 14th century.

5 April 2011

MACHINES

9:00

Dave Lewis – Centre for Next Generation Localisation: Trinity College Dublin
Semantic Model for end-to-end multilingual web content processing

This talk will present a Semantic Model for end-to-end multilingual web content processing flows that encompass content generation, its localisation and its adaptive presentation to users. The Semantic Model is captured in the RDF language in order to both provide semantic annotation of web services and to explore the benefits of using federated triple stores, which form the Linked Open Data cloud that is powering a new range of real world applications. Key applications include the provenance-based Quality Assurance of content localisation and the harvesting and data cleaning of translated web content and terminology needed to train data-driven components such as statistical machine translation and text classifiers.

Alexandra Weissgerber – Software AG
Developing multilingual Web services in agile software teams

Developing multilingual Webservices in agile software teams is a multi-facetted enterprise which comprises various areas that include methodology, governance and localization. We will report on our employment of standards and best practices, particularly where and how they fit or did not fit, and the gaps we have encountered including our strategy to bridge them effectively as well as some of our workarounds.

Andrejs Vasiljevs – Tilde
Bridging technological gap between smaller and larger languages

Small markets, limited language resources, tiny research communities – these are some of the obstacles in development of technologies for smaller languages. In this presentation we will share experience and best practices from EU collaborative projects with a particular focus on acquiring resources and developing machine translation technologies for smaller languages. Novel methods help to collect more training data for statistical MT, involve users in data sharing and MT customization, collect multilingual terminology and adapt MT to terminology and stylistic requirements of particular applications.

Boštjan Pajntar – Joseph Stefan Insitute
Collecting aligned textual corpora from the Hidden Web

With the constant growth of web based content large collections of textual become available. Many if not most professional non-English web sites offer translated webpages to English and other languages of their clients and partners. This are usually professional translation and are abundant. We call this Hidden Web. We intend to present possibilities, problems and best practices for harnessing such aligned textual corpora. Such data can then be efficiently used as a translation memory for example as help for a human translators or as training data for machine translation algorithms.

Gavin Brelstaff – CRS4 Sardinia, Francesca Chessa – University of Sassari
Interactive alignment of Parallel Texts – a cross browser experience

We report our experience test-driving current standards and best-practice related to multilingual Web applications. Following an overview of our pilot demonstrator for the interactive alignment of parallel texts (e.g. poetic translations in/out of Sardinian), we indicate pros and cons of the practical deployment of key standards - including TEI-p5, XML, XSL, UTF-8, CSS2, RESTful-HTTP, XQuery, W3C-range.

Q&A

11:30

Paula Shannon – Lionbridge
Social Media is Global. Now What?

No question about it, companies are embracing social media and working it on a global scale. But the expansion is not without its challenges. Chief among them is how to effectively communicate on multiple platforms, in multiple languages, with a variety of cultural audiences. So how are companies making it happen? In what ways are they using social media globally? What are the emerging best practices for dealing with language and culture on blogs, Twitter, community forums and other platforms?

Maarten de Rijke – University of Amsterdam
Emotions and experiences and the social media

There is little doubt that the web is being fundamentally transformed by social media. The realization that we now live a significant part of our lives online is giving rise to new perspectives on text analytics and to new interaction paradigms. Emotions and experiences are key to communication in social media: recognizing and tracking them in highly dynamic multilingual text streams produced by users around Europe, or even around the globe, is an emerging area for research and innovation. I will illustrate this with a few examples derived from online reputation management and large scale mood tracking.

Gustavo Lucardi – Trusted Translations, Inc.
Nascent Best Practices of Multilingual SEO

The talk will touch, from the perspective of a Language Service Provider (LSP), on how Multilingual Search Engine Optimization (MSEO) is already an essential part of the language Localization process. The presentation will provide an in-depth look at the nascent Best Practices and explain the concepts behind Multilingual Search Engine Optimization.

USERS

Chiara Pacella – Facebook Ireland
Linguistic rules and user-generated content

In social networking websites, a "controlled" component, generated by content creators, must coexist with an "uncontrolled" component, that is generated by the users. Even if the latter is more difficult to control, it is the former that create more challenges in terms of l10n/i18n. The use of a crowdsourcing approach has proven successful for Facebook, but this was achieved thanks to the implementation of standard linguistic rules that are complex and detailed but, at the same time, easily understandable by the actors involved in the translation process.

Ian Truscott - SDL
Customizing the multilingual customer experience – deliver targeted online information based on geography, user preferences, channel and visitor demographics

Users are increasingly using social media and different devices next to the 'traditional' web and offline media. Information that was previously unavailable or inaccessible is today shaping their opinions and buying behaviour. As a result, users' expectations have changed and have raised the bar for any organization that interacts with them. They expect that information is always targeted and relevant to their needs, available in their language and on the device of their choice. The presentation will highlight some of the specific challenges that are emerging as well as demonstrate the technology available to solve them.

Q&A

POLICY

14:30

Jaap van der Meer – TAUS
Perspectives on interoperability and open translation platforms

This presentation will give a summary of the joint TAUS-LISA survey on translation industry interoperability and a report from the recent Standards Summit in Boston (February 28-March 1) as well as perspectives on open translation platforms from TAUS Executive Forums.

Fernando Serván – FAO of the UN
From multilingual documents to multilingual websites: challenges for international organizations with a global mandate

International organizations face many challenges when trying to reach their global audience in as many languages as possible. The Food and Agriculture Organization of the United Nations (FAO) works in six languages (Arabic, Chinese, English, French, Russian and Spanish) to try to have an impact in the agricultural sector of its member countries. The presentation will focus on the need of multilingual support on the Web and will refer to standards and best practices needed . It will cover aspect such as the creation and deployment of multilingual content, the translation needs and possible integration of TM and MT, the availability of CAT tools, etc.

Stelios Piperidis – ILSP - "Athena" RC
On the way to sharing Language Resources: principles, challenges, solutions

This talk will present the basic features of the META-SHARE architecture, the repositories network, and the metadata schema. We will then discuss the principles that META-SHARE uses regarding language resource sharing and the instruments that support them, the membership types along with the privileges and obligations they entail, as well as the legal infrastructure that META-SHARE will employ to achieve its goals. We will conclude by elaborating on potential synergies with neighbouring initiatives and future plans at large.

Q&A

16:00- 16:15
Wrap up & Close

Sponsored by:

Istituto di Informatica e Telematica

Organised by:

World Wide Web Consortium

Consiglio Nazionale delle Ricerche

Istituto di Informatica e Telematica

Istituto di Linguistica Computazionale "A. Zampolli"

European Commission

W3C Italia

World Wide Web Consortium
Ufficio Italiano