

Unicode Localization Data Interoperability TC Overview (ULI)

What's a word? What's a sentence?
Why is this business-relevant?

Christian Lieske, SAP (Walldorf, Germany)

Helena Shih Chapman, IBM (Waltham, Massachusetts, USA)

META  **FORUM 2013**

Context and Overview

META A Network of Excellence forging the Multilingual Europe Technology Alliance

META META-NET META-VISION META-SHARE META-RESEARCH News & Events LT-World Contact

Search Site English

META-FORUM 2013 – Connecting Europe for New Horizons

- > META-FORUM 2013
- > Programme
- > META Exhibition
- > Venue
- > Accommodation
- > Registration
- > Contact

META FORUM 2013

Connecting Europe for New Horizons
September 19/20, 2013

German Federal Ministry of Economics and Technology
Berlin, Germany

META-FORUM 2013 – Connecting Europe for New Horizons is an international multilingual information society, the data value chain and the information at the conference are *Big Data Text Analytics* and *Multilingual Web Services*. The previous editions are META-FORUM 2012, META-FORUM 2011 and META-FORUM 2010.

Important: META-FORUM 2013 takes place at the German Federal Ministry of Economics and Technology. Please be aware of the security check. Please remove any pocket knives, scissors or other sharp objects to provide identification (passport, driver's license etc.). The conference is a bottleneck at the security check, please be at the venue well in advance.

Highlights

Horizon 2020 and Connecting Europe for New Stakeholders: GALA (Globalization and Linguistic Diversity); Council of Europe Committee of Experts on Language Towards a European Language Panel discussion Award Ceremony: META Prize and META Exhibition - industry and research

META-FORUM 2013 will be held jointly by META-NET and the German Multilingual Web-LT, QTLaunchPad and LT Berlin.

META-NET

European Union

German Federal Government

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD

META-FORUM 2013 - META Exhibition

- > META-FORUM 2013
- > Programme
- > META Exhibition
- > Venue
- > Accommodation
- > Registration
- > Contact

META FORUM 2013

Connecting Europe for New Horizons
September 19/20, 2013

German Federal Ministry of Economics and Technology
Berlin, Germany

The 2013 edition of META Exhibition features the following poster presentations and software demonstrations. META Exhibition takes place in Eichenzahl which is next to the Aula where the presentations and discussions of META-FORUM 2013 take place.

1. Tamás Váradi, Hungarian Academy of Sciences (Budapest, Hungary): iTranslate4
2. Robert Grabowski, Acrelix (Berlin, Germany): The ACCEPT project – Helping communities share knowledge across the language barrier
3. Dave Lewis, CNUL (Dublin, Ireland): Use Cases for Exploiting Linked Data in Translation Management
4. Bartelme Metz-Lao, Copenhagen Business School (Denmark): CASMACAT – An interactive post-editing workbench
5. Paul Butelarar, DERI, National University of Ireland (Galway, Ireland): EuroSentiment – Semantically Interoperable Language Resources for Sentiment Analysis
6. Thierry Declercq, DFKI (Saarbrücken, Germany): The TrendMiner project – Multilingual ontology-based analysis of social media
7. Kathrin Eichler, DFKI (Berlin, Germany), Matthias Malsbrock, OIAQ (Berlin, Germany): EXCITEMENT – Exploring Customer Interaction through Textual Establishment
8. Gottfried Herzog, DIN (Berlin, Germany), Laurent Romary, Humboldt-University (Berlin, Germany), Andreas Witt, Institut für Deutsche Sprache (Mannheim, Germany): Standards for Language Resources
9. Balázs Benedek, Easyling (Budapest, Hungary): Easyling.com – website translation proxy
10. Inka Kallias Vihonen, DG Translation, European Commission (Brussels, Belgium): LIND-Web – Joining forces for a stronger language industry and employability
11. Hans Fenstermacher, GALA (Washington D.C., USA), Kim Harris, text & form (Berlin, Germany), Serge Gladkoff, Logrus (Philadelphia, USA): GALA's LT Advisor – Year sources for language technology specs and reviews
12. Serge Gladkoff, Logrus (Philadelphia, USA): The WTC5 project
13. Kurt Eberle and Friederike Weissenfels, Lingemio (Heidelberg, Germany): On-the-fly extraction of dictionary information for MT services from bilingual corpora
14. Matthias Wendt (Neofonie, Germany), Feiyu Xu (DFKI, Germany): Semantic search technologies for medical information extraction and question answering
15. Pavel Rychlý, NLP Centre, Masaryk University (Brno, Czech Republic), Miloš Jakubíček, Lexical Computing Ltd. (Brighton, UK): Crapping 70+ billion word corpora in SketchEngine
16. Stella Pateraki and Juhl Bakajani, I.C. Athina /TISP (Athens, Greece), Christian Spurr, DFKI (Saarbrücken, Germany), Khalid Chouki and Olivier Hamon (ELDA, France): META-SHARE, the open language resource exchange facility
17. Jiří Wichter, Pataman (Prague, Czech Republic): PMSE – A Generic Framework for Corpus Processing, Analysis and Visualization
18. Bastian Enners, Doris Langenberg, Benjamin Luetke, Nancy Radloff and Daniel Reib, Plunet (Berlin, Germany): Plunet Business Manager – it's more than a software – it's your business! Efficient Translation Project and Business Management in one flexible solution
19. Xianfeng Cheng, Speechoscan (Beijing, China): ASR speech corpus building in European languages
20. Ines Liebscher, textform (Berlin, Germany): Multilingual Terminology Workflow Using Rule-Based MT for High-Quality Translation Scenarios
21. Michael R. Alvers, Transinsight (Dresden, Germany): Enterprise Semantic Intelligence (ESI 8): a search technology for enterprises
22. Holmer Hansen, Torsten Küllas, Johannes Kirschnick, Database Systems and Information Management Group, Technical University Berlin (Germany): MIA – Big Data Analytics
23. Jochen Adams, Technical University Berlin (Germany): Data Supply Chains for European Data Pools with Stratosphere
24. Allan Hanbury, Vienna University of Technology (Austria), Jan Hajič, Charles University in Prague (Czech Republic): Khresmoi – Multilingual Metadata Interoperability
25. Christian Lieske, SAP (Walldorf, Germany) and Helena Shih Chapman, IBM (Waltham, Massachusetts, USA): Unicode Localization Data Interoperability (ULI) Technical Committee Overview
26. Christian Lieske, SAP (Walldorf, Germany) and Helena Shih Chapman, IBM (Waltham, Massachusetts, USA): Unicode Localization Data Interoperability (ULI) Technical Committee Overview
27. Philippe Lacroix, University of Technology of Troyes (France): Multilingual and Precise Web Translation – the TraduXio Project
28. Lutz Rilling, Vocoy (Berlin, Germany): Next Generation Language Guide

The Unicode Localization Interoperability Technical Committee (ULI-TC) was established in 2011 with the goal of helping to ensure interoperable data interchange of critical localization-related assets. ULI's work is relevant to speech/natural language processing, analytics tokenization etc. including translation memories, segmentation rules, and more. What ULI is building forms the foundation of many other downstream technologies: memory interchange, speech/natural language processing, analytics tokenization etc.

META-FORUM 2013 – Connecting Europe for New Horizons

Christian Lieske, SAP (Walldorf, Germany), IBM (Waltham, Massachusetts, USA): Unicode Localization Data Interoperability (ULI) Technical Committee Overview

Unicode & Segmentation (1/3)

- More than a character repertoire – an **ecosystem**, a **stack of standards**
- Parts of the ecosystem are related to “segmentation” questions such as “How can text entities such as **sentences** be broken down into sub-entities such as **words**?”
- Segmentation is important for **business analytics** and **translation**...

Unicode & Segmentation (2/3)

Most prominent members of the Unicode ecosystem related to segmentation:

- Unicode Text Segmentation report

TR#29 <http://www.unicode.org/reports/tr29>

- Unicode Line Breaking Algorithm

TR#14 <http://www.unicode.org/reports/tr14>

- Common Locale Data Repository

CLDR; see <http://cldr.unicode.org>

Unicode & Segmentation (3/3)

Comprehensive support for Unicode is provided by the International Components for Unicode (ICU, www.icu-project.org), a **software library** used in many applications.

ULI Credo

If Unicode and its “citizens” CLDR, and ICU get segmentation right, many applications get text processing right:

- Business analytics
- Speech/natural language processing
- Memory interchange
- Sorting
- Searching



ULI Scope & Objectives

- **Gather requirements** for core and extension of the standards in the area of text segmentation and content memory
- **Establish core specification scope**, extension domain, and reference implementation to improve the usefulness of existing standards
- Create a repository of reference user profile and scenarios to **demonstrate interoperability** across desired standards
- Provide **consistent interpretation of the specification**, extension and profiles

ULI Setup

Logistics

- Meet once a month by telephone
- Regular participation by IBM, Microsoft, Yahoo, Google, SAP, Globalization and Localization Association (GALA), and XML Localization Interchange File Format Technical Committee (XLIFF TC)

Challenges

- Need more translation tool vendor involvement
- Solicit additional participation from key industry conferences

Open for participation

- Active participation is expected
- Need to be a member to attend meetings regularly
- For details, see [TC Procedure](#) on Unicode site

ULI 2012

Internal agreement on plain text content boundary joining and separate best practices:

- Leveraging TR#29
- Agreed syntax for referencing CLDR elements (XPath to the CLDR parent element level; initially vetted English, German, Russian, and Spanish – see <http://unicode.org/uli/trac/browser/trunk/abbrs>)
- Demoed behavior of updated ULI input (see <http://demo.icu-project.org/icu-bin/icusegments>)

ULI 2013/2014

- Draft implementation to demonstrate ULI progress
- CLDR and ICU contribution integration:
 - Initial ULI input for sentence level segmentation submitted to CLDR 24 due September 15, 2013 (see <http://cldr.unicode.org/index/downloads/cldr-24>)
 - Plugin implementation to ICU in progress for ICU 52 due October 2013 (see <http://site.icu-project.org/download>)
- Open source Computer-Assisted Translation integration in 2014 (ongoing evaluation of ICU implementation, based on ULI input into OpenTM2, see <http://www.opentm2.org>)