LIDER: FP7 – 610782

Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe

Deliverable number **D4.6**

Deliverable title **Report of the 2nd Roadmapping Workshop**

Main Authors **Felix Sasaki**

Grant Agreement number	610782
Project ref. no	FP7-610782
Project acronym	LIDER
Project full name	Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe
Starting date (dur.)	1/11/2013 (24 months)
Ending date	31/10/2015
Project website	http://www.lider-project.eu/

Coordinator	Asunción Gómez-Pérez
Address	Campus de Montegancedo sn. 28660 Boadilla del Monte, Madrid, Spain
Reply to	asun@fi.upm.es
Phone	+34-91-336-7417

Fax	+34-91-3524819

Document Identifier	D4.6
Class Deliverable	LIDER EU-ICT-2013-610782
Version	1.1
Document due date	31 July 2014
Submitted	14 August 2014
Responsible	W3C/ERCIM
Reply to	fsasaki@w3.org
Document status	draft
Nature	O(Other)
Dissemination level	PU(Public)
WP/Task responsible(s)	Felix Sasaki, DFKI / W3C Fellow
Contributors	-
Distribution List	Consortium Partners
Reviewers	Reviewed by the project consortium
Document Location	http://lider-project.eu/?q=doc/deliverables

Executive Summary

This document summarizes contains three reports:

- The second LIDER roadmapping workshop[1], held 8-9 May with 45 registered participants and in alignment with the MultilingualWeb workshop in Madrid
- The MultilingualWeb workshop report[2]. The workshop was held 7-8 May in Madrid and had 110 registered participants.
- The third LIDER roadmapping workshop[3], held 4 June with 40 participants in alignment with Localization World in Dublin.

In addition, the LIDER project has started to gather key outcomes of roadmapping activities. The current state of these outcomes is available at

https://www.w3.org/community/ld4lt/wiki/Lider_roadmapping_activities

This page will be kept up to date during the duration of the project.

[1] See the online version oft the report at http://www.multilingualweb.eu/documents/2014-madrid-workshop/2014-madrid-workshop-report#lider_roadmapping

[2] See the online version oft the report at http://www.multilingualweb.eu/documents/2014-madrid-workshop/2014-madrid-workshop-report

[3] See the online version oft the report at https://www.w3.org/community/ld4lt/wiki/LD4LT_Roadmapping_Workshop_Dublin_June_2014

Document Information

IST Project Number	FP7-610782	Acronym	LIDER
Full Title	Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe		
Project URL	http://www.lider-project.eu/		
Document URL	http://lider-project.eu/?q=doc/deliverables		
EU Project Officer	Susan Fraser		

Deliverable	Number	D4.6	Title	Report of the 2nd Roadmapping Workshop
Workpackage	Number	4	Title	Community building and dissemination

Date of Delivery	Contractual	31 June 2014	Actual	14 August 2014
Status	version 1.1		final ■	
Nature	prototype □ report □ dissemination ■			
Dissemination level	public ■ consortium □			

Authors (Partner)	Felix Sasaki, DFKI / W3C Fellow			
Responsible Author	Name	Felix Sasaki	E-mail	fsasaki@w3.org
	Partner	DFKI / W3C Fellow	Phone	+49-30-23895-1807

Abstract (for dissemination)	
Keywords	LIDER, roadmapping workshop, report

Version	Modification(s)	Date	Author(s)
01	First Draft	6/08/14	Felix Sasaki, DFKI / W3C Fellow
02	Revision	11/08/14	Asunción Gómez-Pérez, UPM
02	Final version	14/08/14	Felix Sasaki, DFKI / W3C Fellow

Project Consortium Information

Participants		Contact
Universidad Politécnica de Madrid		Asunción Gómez-Pérez Email: asun@fi.upm.es
The Provost, Fellows, Foundation Scholars & The Other Members of Board of The College of the Holy & Undivided Trinity of Queen Elizabeth near Dublin (Trinity College Dubl, Ireland)		David Lewis Email: dave.lewis@cs.tcd.ie
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)		Felix Sasaki Email: felix.sasaki@dfki.de
National University of Ireland, Galway (NUI Galway, Ireland)		Paul Buitelaar Email: paul.buitelaar@deri.org
Institut für Angewandte Informatik EV (INFAI, Germany)		Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Universität Bielefeld (UNIBI, Germany)		Philipp Cimiano Email: cimiano@cit-ec.uni-bielefeld.de
Universita degli Studi di Roma La Sapienza (UNIVERSITA DEGLI STU, Italy)		Roberto Navigli Email: navigli@di.uniroma1.it
GEIE ERCIM (ERCIM, France)		Felix Sasaki Email: fsasaki@w3.org

Table of Contents

1 Introduction

This document summarizes contains three reports:

- Section two contains the report for the second LIDER roadmapping workshop, held 8-9 May in alignment with the MultilingualWeb workshop in Madrid
- Section three contains the report for the third LIDER roadmapping workshop, held 4 June in alignment with Localization World in Dublin.
- The third part of this document is a PDF version generated from the online MultlingualWeb workshop report. The workshop was held 7-8 May in Madrid. The report is an HTML, multimedia document. Hence it is recommended to read the report online at

http://www.multilingualweb.eu/documents/2014-madrid-workshop/2014-madrid-workshop-report

The LIDER project has started to gather key outcomes of roadmapping activities. The current state of these outcomes is available at

https://www.w3.org/community/ld4lt/wiki/Lider_roadmapping_activities

and will be kept up to date during the duration of the project.

2 Report: 2nd LIDER Roadmapping Workshop, Madrid 8-9 May 2014

2.1 *Introduction*

This report summarizes the LD4LT / LIDER roadmapping workshop, held with 45 registered participants 8-9 May 2014, in alignment with the MultilingualWeb 2014 workshop, which had 110 registered participants.
The LD4LT / LIDER roadmapping workshop targeted feedback from the Web community at large on aspects of linked data and language technology. Two specific focus topics were the role of Wikipedia based data sources and the conversion of existing resources into linked data. Below is a summary of each contribution and a list of key findings.
The workshop consisted of seven parts: a keynote presentation from Seth Grimes, two panels on multilingual aspects of Wikipedia, a use case and requirements session, and three sessions on linked data and language technologies.

2.2 *Contributions*

Text Analytics, Applied. In his keynote presentation on Text Analytics Applied, Seth Grimes introduced approaches and usage scenarios of text analytics. Masses of unstructured content in many different forms (Web pages, emails, multimedia content, social media etc.) contains named entities, facts, relationships, sentiment information etc. Text analytics turns these into structured information for various usage scenarios. The presentation and the related report on text analytics provide information on application areas, the type of information actually being analysed in industry, or current language coverage. Important points for LIDER roadmapping were:

- Text analytics generates the structured information for bridging search, business intelligence and applications.
- Key real-life application areas are: business intelligence, life sciences, media & publishing, voice of the customer, marketing, or public administration and policy making.
- Users of text analytics tools need: adaptability to their content domain, customization (e.g. import of taxonomies), flexible input & output formats and processing mechanisms (API, offline, ...).
- Sentiment resolution is an important functionality for many usage scenarios.
- When deciding on solutions, users take capacity (volume, performance, latency) and cost into account.
- Multilingual text analytics is an area that so far has not seen a lot of activity in industry.

Using Wikipedia for multilingual web content analytics. In this panel session, Pau Giner, Amir Aharoni and Alolita Sharma provided details about the Wikipedia translation infrastructure. So far users only have indicators, but no explicit provenance information about what article is a direct translation from other languages. There is also no strict translation workflow. This is due to the nature of the Wikipedia community, including Wikipedia editors, translators and new users who do content creation via translation. Significant points of the session are summarized below.

- For its content translation tool, Wikipedia is looking into automatic translation tools, allowing the user to translate per paragraph and revise the result.

- Handling feedback from users for the tool development is a challenge since about 200 communities have to be taken into account.
- Wikipedia based machine translation tooling could help to quickly create domain specific MT.
- The multilingual infrastructure of Wikipedia could be the basis for new research topics like comparing content across cultures and in this way cultures themselves.

Growing Wikipedia editing with intelligent multi-language suggestion lists for article translation as well as other techniques and tools. This session started with Runa Bhattacharjee, Pau Giner and Santhosh Thottingal on a panel. It was highly interactive, which is reflected by the summary of key points below.

- Information on translation quality could help translators in Wikipedia. Such information is being defined in the QTLaunchPad project, see the presentation at the MultilingualWeb workshop from Arle Lommel.
- Various types of background information can help translators in several ways: providing translation suggestions, disambiguate terms, autocompletion etc.
- Data models for structured information in Wikipedia, e.g. Wikidata, do not rely on the linked data, that is RDF technology stack. But conversions to and from linked data are possible
- One challenge is the integration of resources like Wikidata or others that have been discussed at the MultilingualWeb workshop, like BabelNet or DBpedia, as discussed in the presentation from Roberto Navigli and Martin Brümmer et al.

Gathering of use case and requirements.
To further investigate and understand the use cases and requirements of LOD the participants were asked to participate in a post-it sessions. Outcomes of a similar session from the previous roadmapping workshop are described in section 4 of D1.1.1. This session involved two sets of post-its, one related to use cases and one related to requirements. The participants provided approx. 40 different use cases and requirements. Based on the main theme being LOD most use cases and requirements reflected core LOD topics such as supporting MT through access to large (annotated) data sources.
However, interestingly larger trends around Linguistic data were identified and can be summarized as 'Access', 'Connectivity', 'Contribute' and 'Understand'.
'Access' relates to several comments indicating the need for more open APIs and data sources that can be used to support and augment existing data sets and applications. 'Connectivity', a core theme in LOD, was highlighted by the need for more RDF based data sources. 'Contribute' was indicated by comments seeking a stronger support and contribution by communities and community driven projects (e.g. Wikimedia and Wordnet). Finally, 'Understand' relates to needs identified in the area of sentiment and content analysis.
It can be concluded that most use cases and requirements related to providing access to more structured/annotated data sources and allowing simple connectivity via APIs and standards.

Initiatives and community groups related to linked data and language technologies. In this session key persons provided an overview of various groups. The groups presented are LIDER (presented by Asunción Gómez-Pérez), FALCON (presented by Dave Lewis), the LD4LT Community group (also presented by Dave Lewis), the OntoLex Community group (presented by Philipp Cimiano), the Best

Practices for Multilingual Linked Open Data (BPMLOD) Community Group (presented by Jorge Gracia) and the Open Linguistics Working Group of the OKF (presented by Christian Chiarcos). Key points of the discussion were:

- One has to be careful in classifying resources. Some are language resources (e.g. lexica), others are general resources like Wikipedia / DBpedia / Wikidata. All of these can be relevant for linguistic processing, but they are different in nature.
- An example of this aspect is Wikipedia. It is a knowledge based often used in natural language processing, but not a linguistic resource.
- A (diagram of an) linguistic linked data cloud needs to clearly distinguish between different types of resources, since it is an entry point for potentially interested people or institutions.
- The quality of resources is sometimes hard to evaluate. Looking at scientific publications can help, e.g. a resource mentioned frequently may be of interest.

Data and Metadata of Language Resources as Linked Data on the Web. The aim of this session was to discuss general approaches and concrete examples of how to represent language resource metadata and the resources themselves as linked data. Christian Chiarcos discussed ISO TC 37 originated standards like LMF or GrAF and RDF as a representation model. He when discussed various parts of the linguistic linked data cloud (corpora, lexical resources, term bases) and use cases and certain challenges for representing these as linked data.
Philipp Cimiano presented several use cases in the form of queries on linguistic resources, and presented a proposal for a linked data based architecture to realize these use cases. Interoperability of metadata is a key challenge to make this vision happen. The LD4LT community group is working into this direction. One key metadata vocabulary is META-SHARE. Stelios Piperidis provided background on META-SHARE and its role in the planning of a new machine translation initiative QT21, see the presentation from Hans Uszkoreit at the MultilingualWeb workshop for details. Marta Villegas introduced the current state of mapping the META-SHARE XML Schema to RDF.
Roberto Navigli provided detailed technical aspects of BabelNet, complementing his general overview at the MultilingualWeb workshop. BabelNet relies on existing linked data vocabularies like Lemon, SKOS or LexInfo 2.0. RDF modeling issues arise when linking linguistic information to general knowledge e.g. in Wikipedia. Thierry Declerck showed how dialectal dictionairies can be represented as RDF and integrated with general linked data information. The encoding of textual information associated with entries and senses (e.g. examples) is a challenge.

Multilingual Corpus transformation into Linked Data. Here, the RDF based NIF format can help. Martin Brümmer introduced NIF. It aims to foster interoperability between natural language processing tools, language resources and annotations. In this way, NIF can serve as a pivot format for corpora. Roberto Navigli presented a concrete corpus that shows the application of NIF and ITS 2.0 information: an RDF version of the XML-based MASC corpus. Felix Sasaki demonstrated an application of NIF integrating both linguistic information and localization workflow metadata.

The above presentations clearly demonstrated the feasibility of NIF based corpus creation and processing. Laurette Pretorius closed the workshop discussing opportunities and challenges for representing information related to under-resourced languages of Southern Africa as linked data. The processes of publication, discovery, sharing and consuming of language resources could greatly benefit from linked data. To

make this happen, proper tool infrastructure needs to be in place and best practices on how to work with linguistic linked data need to be made available.

2.3 *Key points of the Workshop*

The workshop had a huge variety of topics and sessions, including an interactive session on gathering requirements and use cases for linked data. A summary including key points of the workshop in general is below.

- About text analytics:
  - Users of text analytics tools need: adaptability to their content domain, customization (e.g. import of taxonomies), flexible input & output formats and processing mechanisms (API, offline, ...).
  - Sentiment resolution is an important functionality for many text analytics usage scenarios.
  - When deciding on solutions, users take capacity (volume, performance, latency) and cost into account.
  - Multilingual text analytics is an area that so far has not seen a lot of activity in industry.
- About multilingual aspects of Wikipedia:
  - Various types of background information also from Wikipedia can help translators.
  - Data models for structured information in Wikipedia, e.g. Wikidata, do not rely on the linked data,but conversions to and from linked data are possible.
  - One challenge is the integration of resources like Wikidata, BabelNet or DBpedia.
- On language resources and language resource metadata:
  - One has to be careful in classifying resources, e.g. language resources (e.g. lexica) versus general knowledge bases like Wikipedia
  - A (diagram of an) linguistic linked data cloud needs to clearly distinguish between different types of resources.
  - The quality of language resources is sometimes hard to evaluate.
  - Interoperabilty of language resource metadata is a key prerequisite for working with the linguistic linked data cloud.
- On multilingual corpus transformation:
  - NIF can serve as a pivot format for linked data based corpus representation and processing.
  - Several presentations demonstrated the feasibility of a NIF based approach in different usage scenarios, like RDF representation of existing, XML-based corpora, or integration of linguistic information with localization process metadata.
  - Tooling should be made available to work with linguistic linked data; best practices are needed as well.

Many use cases and requirements discussed at the workshop related to providing access to more structured/annotated data sources and allowing simple connectivity via APIs and standards.

3 Report: 3rd LIDER Roadmapping Workshop, Dublin 4 June 2014

3.1 *Introduction*

This report summarizes the LD4LT / LIDER roadmapping workshop, held 4th June 2014 as part of FEISGILTT workshop with 40 participants. The workshop was co-located with Localization World Dublin 2014. See more information about the workshop program.
The workshop targeted feedback from the localization community on linked data and language technology. Below is a summary of each contribution and a list of key findings.

3.2 *Contributions*

Welcome and Introduction
Dave Lewis (TCD) and Felix Sasaki (DFKI / W3C Fellow) introduced the workshop and its goals. The localization industry so far has not a lot of experience with linked data in real use cases. The workshop brings the right people from industry to find out: what problems could be solved by using linked data? What use cases are needed? What hinders the adoption of linked data?
Dave and Felix also gave an introduction to the umbrella LIDER project. LIDER works on the basis for creating a Linguistic Linked Data cloud, which can support content analytics tasks of unstructured multilingual cross-media content. Localization is an important usage scenario of LIDER, hence this roadmapping workshop.

Phil Richie: Translation Quality and RDF
Phil Ritchie (VistaTEC) reported on experience with two technologies: ITS 2.0 and NIF. At VistaTEC, a localization workflow using quality related ITS 2.0 information has been created. The Ocelot tool can be used to visualize such information and to correct translations. Reyling on NIF, VisaTEC produced linked data representations of translations. Here the benefit of linked data lies in data integration: various information related to a translation quality and other aspects can be freely integrated.
For localization companies, localization has to happen faster and faster, and prices are going down. So the companies are looking for added value, that is beyond the actual translation. The concept of **intelligent content** could convey such value. But the creation of such content requires new skills for translators and authors.
Using linked data in localization may only happen if companies can do this without opening the data. Many data sources are the assets of the company. Also, licensing mechanisms for linked data are a key aspect to make linked data for localization happen.

Yves Savourel: Linked Data in Translation Kits
Yves Savourel (ENLASO) gave a demonstration with the okapi toolkit, using linked data sources in a localization workflow. Linked data sources can help to give several types of information to the translator: general context information, definitions, disambiguation, translation examples etc. A usability challenge is: how to present the information to the translator so that it really is helpful, speeds up the localization process and leads to better quality? The issue can be summarized as: "Too much information is no information".
A technical challenge is overlap of information. Overlapping is no issue with NIF, that is in the RDF representation. But most of the localization tools work with XML data, and the

standardized formats (XLIFF, TMX, TBX, ...) in localization do not provide mechanisms to represent overlapping information.
The discussion after the presentation provided additional points. If the quality of context information is unclear it may be rather a burden than a help. Always up to date information is needed. The forehand mentioned technical challenge (linked data not natively supported in localization formats) could be resolved by creating standardized JSON representations on top of these formats.

David Lewis: Turning Localisation Workflow Data to Linked Data
Dave Lewis introduced the FALCON project. The aim of FALCON is to provide the basis for a "localization Web". Here, resources used during localization (e.g. terms, translations) become linkable resources. Linkable Metadata in localization workflows then provides added value, compared to current "silo" approach: today, data held used in localization is stored and processed often in a proprietary and non-interlinkable manner.
A localization Web can help to leverage automatic language processing. For example, linked information can be used to leverage machine translation training or text analytics tasks. Core use cases in FALCON are:

- source content internationalisation, with term extraction and translation discovery.
- machine translation, producing consistent translations of terms and including discovery of parallel text for training.
- translation and post-editing, including term definitions from open encyclopaedic data like Wikipedia and concordancing over a global translation memory.

Alan Melby: Linport and RDF
Alan Melby (Brigham Young University) reported on the Linport project. Linport is developing a format to package translation materials. The package will be self-contained, platform or translation tool independent, and it will come with basic operations like splitting or merging packages.
Linport defines an XML-based format. So far the group has not looked into using RDF. The discussion at the workshop around Linport and RDF did not lead to conrete steps in this direction.
Alan also reported on quality related efforts, namely MQM and DQF. Harmonization efforts between these are underway and a joint framework would be highly desirable.

Andrejs Vasiljevs: Terminology resources in the ecosystem of linked data services for language professionals
Andrejs Vasiljevs (Tilde) presented the TaaS project. TaaS provides a cloud-based platform for instant access to up-to-date terms, user participation in term acquisition and sharing, and terminology resources reuse. The platform allows to automatically extract term candidates, to acquire translation equivalents or to clean up user provided resources.
TaaS also comes with import and export APIs. The export into linked data is one work area undertaken in cooperation with the LIDER project. The discussion after the presentation it became clear that linked data representation of terminology information is a topic of huge interest. The discussion fed into work of the LD4LT group, see the related issue Terminology and linked data.

Ioannis Iakovidis: TermWeb and The Localisation Web

Ioannis Iakovidis ([Interverbum Technology]) introduced TermWeb, their SaaS Terminology Management Solution. He described how its concept based approach to terminology management and its implementation as a web application made integration with lexical-conceptual resources captured as linked data a natural next step for the evolution of their product.
Integration with client's content management and localization workflows is key in deploying TermWeb. By participating in the FALCON, Interverbum aims to reap benefits of; broader sharing of term based; linking into public terminological resources, e.g. BableNet; providing links for auditing and providing quality assessment on term bases and leveraging term translation in machine translation.

Víctor Rodríguez Doncel: Towards high quality, industry-ready Linguistic Linked Licensed Data
Víctor Rodríguez Doncel (UPM) touched upon a topic that was of huge interest in many discussions: licensing and linked data. For the localization community, fully open linked data may not be of high relevance. Hence, a licensing mechanism is deeply needed to foster linked data adoption.
Different licensing models have an impact on business opportunities for linked data. The Open Digital Rights Language provides a framework for expressing such models in a machine readable manner. For localization, "Licensed Linguistic Linked Data (3LD)" may be of most interest. Here, different licensing models can be used together, including totally open and restrictive licenses, or completely closed datasets.

3.3 *Key points of the Workshop*

The workshop closed with an interactive session on gathering requirements and use cases for linked data. A summary including key points of the workshop in general is below.

- Text analytics
  - There is a need for a common API to text analysis services, e.g. Babelnet, DBPedia spotlight, wikidata, Yahoo! content Annotation.
  - One needs to support attribution of source of lexical/terminologycal data and meta-data - especially when using aggregtion services such as Babelnet.
  - Resources need to have live updates to assure constant improvement.
  - Users need a mechanism to feed back corrections to annotation service, also into the underlying resources.
  - JSON can be used to provide annotation meta-data, e.g. as part of common API or as payload across different APIs.
  - One needs to be able to indicate the relevance of an annotation, e.g. confidence scores and their consistent interpretations.
  - Understanding context is key to assessing quality of annotations.
  - A stand-off annotation mechanism is needed to deal with annotation overlap. NIF could be a solution.
  - What type of CAT tool support is needed, e.g.: access to definitions, access to usage examples, use in predictive typing.
- Licensing metadata
  - Licensing information needs to be integrated with the actual data.
  - One needs to be able to automatically compound different license terms to enable understanding at point of use as end of value chain.
- Localisation project metadata

    - The relationship between efforts like Linport's STS, EC's MED and RDF should be made clear.
- Terminology information and RDF
    - There is no standard mapping of the TBX format to RDF.
    - How should terminology information be incorporated with text analysis information and an related API?
    - How should one integrate open lexicons with closed corporate term bases?
- Bitext
    - One could expose bitext (= aligned text of a source and one or several translations) as linked data, as an alternative to TMX.

4 Report: MultilingualWeb workshop, Madrid, 7-8 May 2014

The report of the “MultilingualWeb workshop, Madrid, 7-8 May 2014”, is available online at

http://www.multilingualweb.eu/documents/2014-madrid-workshop/2014-madrid-workshop-report

Below is a PDF version of the document.

W3C Workshop Report: New Horizons for the Multilingual Web

7 – 8 May 2014, Madrid

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 the LIDER project.

On 07-08 May 2014 the W3C ran the seventh workshop in the series. The theme of the workshop was “New Horizons for the Multilingual Web”. The Madrid workshop was hosted by the Universidad Politécnica de Madrid. Félix Pérez Martínez, Director de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la UPM (ETSIT UPM), and Victor Robles Forcada, Director de la Escuela Técnica Superior de Ingenieros Informáticos de la UPM (ETSIINF UPM) gave a brief welcome address.

As with the previous workshops, this event 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 had more than 110 registered participants and featured one and a half days of talks plus a half day open space discussion. A specific focus was on multilingual linked data, a key topic in a dedicated LIDER roadmapping workshop. In the open space slot, participants suggested ideas for discussion groups and then split into two groups, one of which was the LIDER roadmapping workshop. This workshop had 45 registered participants.

We were able to stream the content live on the internet and we also recorded the presentations, which are available on the Web using the MultilingualWeb Youtube channel. We also once more made available live IRC scribing to help people follow the workshop remotely and so assist participants in the workshop itself.

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 summary of each talk accompanied by a selection of key messages in bullet list form. Links are also provided to the IRC transcript (taken by scribes during the meeting), video recordings of the talk, 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.

The creation of this report was supported by the European Commission through the Seventh Framework Programme (FP7), Grant Agreement No. 610782: the LIDER project. The MultilingualWeb workshop series is being supported by LIDER and has been supported by the Thematic Network MultillingualWeb, Grant Agreement No. 250500, and by the LT-Web project, Grant Agreement No. 287815.

Contents:

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.

Workshop sponsors

VERISIGN

Lionbridge

Supported by LIDER

Several presentations focused on multilingual aspects of **Wikipedia**. This started with the keynote speaker, Alolita Sharma, who gave a broad overview of the topic. It continued with Pau Giner, David Chan and Santhosh Thottingal. They provided details on technical developments. Currently Wikipedia supports 287 languages, and there is a huge imbalance in terms of coverage. The Wikimedia foundation and the Wikipedia community at large are working on changing the situation. Translation tooling infrastructure is just one aspect of this development. Resolving this imbalance is an important task in the next years. Otherwise, as Georg Rehm pointed out in his presentation, some languages may face even digital extinction on the Web.

Another main topic of the workshop was **structured, multilingual data sources**. Some of such data sources are closely related to Wikipedia, like Wikidata or DBpedia. The latter was the topic of Martin Brümmer, Mariano Rico and Marco Fossati. They introduced the overall DBpedia infrastructure and aspects of DBpedia language chapters in Italy and Spain. Search engine provides have started the Schema.org effort that defines structured data items for Web content authors. Charles McCathie Nevile reported on multilingual aspects of Schema.org. Roberto Navigli presented BabelNet, a huge multilingual, lexical resource, and its related multilingual disambiguation and entity linking service, Babelfy.

Structured, multilingual data sources are also the main topic of the LIDER project and the LIDER roadmapping workshop. Asunción Gómez-Pérez introduced the general objectives of the project: to create the basis for a Linguistic Linked Data cloud. One usage scenario for this data is content analytics of unstructured multilingual cross-media content. But there are many others: Marion Shaw emphasized that structured content is an opportunity to speed up the localization process. This is urgently needed for example because more and more new devices have to be covered in localization.

Before considering such applications of structured data, one needs to consider fundamental questions like licensing, see the presentation from Victor Rodríguez Doncel. Users of such data sources also need guidance on how to deal with them, or how to convert existing data sources to linked data. This was the topic of Jorge Gracia and José Emilio Labra. The provided the latest state of work undertaken in the W3C Best Practices for Multilingual Linked Open Data Community Group. The LD4LT Community group explores usage scenarios and requirements for multilingual linked data and is also a forum for discussing the conversion of existing data into linked data. Tatiana Gornostay covered this aspect in her presentation on the online portal “Terminology as a Service (TaaS)” by discussing the conversion of terminology resources.

Another focus topic of the workshop was standards related to **multilingual workflows**. That encompasses the localization workflow, see the presentations from Jan Nelson on XLIFF and its application in the Multilingual App Toolkit Version 3.0 from Microsoft, or Joachim Schurig on usage of XLIFF in Lionbridge, and David Filip on the latest state of XLIFF 2.0. Closely related is ITS 2.0, see the presentation from Dave Lewis. ITS 2.0 conveys metadata (so-called “data categories”) for

content creation, localization and other parts of multilingual workflows. Such metadata can improve the quality of post-editing after machine translation, see the presentation from Celia Rico for details. Quality related information was the focus of Arle Lommel, who introduced the Multidimensional Quality Metrics (MQM).

Crowdsourced or curated, multilingual data sources in standards based multilingual workflows: this is a future vision that clearly shows new horizons for the multilingual Web. The presentation from Hans Uszkoreit showed a vision in a different but related area: high quality machine translation. But already now the number of **multilingual, Web related applications** is growing continuously. Feiyu Xu introduced the technology behind the “Yocoy Language Guide”. Fernando Servan explained how the Food and Agriculture Organization of United Nations (FAO) removed language barriers on their website at least partially, relying on machine translation. Seth Grimes described usage scenarios and methods of sentiment analysis on the Web. And Alexander O'Connor introduced usage of multilingual technologies and related standards in the realm of digital humanities.

Depending on the domain in question, **multilingual Web application areas have various needs**. Gemma Miralles presented the scenario of tourism and Web sites in several languages. Here, translation has to take aspects like common keywords in user search into account. The keywords may depend on the users region, hence this is an important aspect of translation. Rob Zomerdijk argued in a similar manner and explained how discovery of content depends on understanding the target audience.

Although many presentations at the workshop were related to technology, it became clear that **the user can and must have an impact** so that the multilingual Web can develop fully. In the presentation from Pedro L. Díez Orzas, the users of multilingual technologies are small and medium sized localization companies. If multilingual technologies are too expensive, they cannot keep a competitive position on the market. Don Hollander and Dennis Tan focused on IDNs (Internationalized Domain Names). IDNs have a clear benefit for all users, enabling them to create Web addresses in their own script. So far IDNs are not widely recognised by users. A holistic view like in the MultilingualWeb community is needed, to move adoption of IDNs in the whole Web tooling ecosystem forward.

Richard Ishida made a call for action in his presentation, by putting certain users in the driver seat. **International layout requirements** can only be integrated into core Web technologies like CSS or HTML if layout experts from various layout communities provide input. A success story in this area is the Requirements for Japanese Text Layout document. Similar documents are currently being created for other layout traditions.

The last block in the day consisted of two parts. In the open space discussion session, the topics of data formats for the multilingual web and IDNs came up again. In addition, the future of the MultilingualWeb workshop series was discussed. In the LIDER roadmapping workshop, discussion on multilingual content analytics and Wikipedia were continued. An additional topic was the transformation of existing language resources into linked data.

Welcome session & Keynote talk


Félix Pérez Martínez, Director de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la UPM (ETSIT UPM), and **Victor Robles Forcada**, Director de la Escuela Técnica Superior de Ingenieros Informáticos de la UPM (ETSIINF UPM), welcomed the participants and emphasized the importance of the workshop theme. This was followed by a brief welcome from **Arle Lommel**, co-chair of the Workshop.

Related links: IRC

The keynote was given by **Alolita Sharma** (Wikimedia Foundation). She spoke about “Multilingual User Generated Content at Wikipedia scale”. Wikipedia currently covers 287 languages, but most of the content is available for a rather small number of languages. Content consumption practices are also very different in between regions, with Europe on the top, followed by Asia-Pacific and US/Canada. Wikipedia is working on changing this situation, with new, cross-lingual infrastructure like Wikidata and new means to support translation of Wiki content. Other significant remarks:

Related links: Slides • IRC • Video

> A large number of Wikipedia language versions include less than 100000 articles.
> For some regions and countries it is crucial to provide access to Wikipedia via many different devices, taking especially the mobile Web into account. For example in Japan it is common to access the Web every day with four different type of devices.
> A community with passion about Wikipedia is key to increase content creation and consumption.

Developers session

The developers session was chaired by **Philipp Cimiano**, Bielefeld University.





Pau Giner, **David Chan** and **Santhosh Thottingal** (Wikimedia Foundation) began the session with their presentation on “Best Practices on the Design of Translation”. They introduced the efforts of the language engineering team from the Wikimedia foundation. The goal is to build open source tools to facilitate the translation of content when creating new articles. This should help to spread high quality content across languages quickly. Other significant remarks:

Related links: Slides • IRC • Video

- Translation tooling in Wikipedia includes work on a great variety of tools, including machine translation or dictionaries.
- Wikipedia also involves semantic knowledge sources into its cross-lingual infrastructure, like Wikidata.
- The tool infrastructure should help to create high quality translations, otherwise it will not be accepted by the Wikipedia community.

Feiyu Xu (DFKI, Co-Founder of Yocoy) presented on “Always Correct Translation for Mobile Conversational Communication”. The usage scenario for the “Yocoy Language Guide” is to provide always correct translations for travelers. The mobile application focuses on face-to-face communication and situation-based dialogues. Other significant remarks:

Related links: Slides • IRC • Video

- The application uses a new technology called ACT (Always Correct Translation).
- Conversation-based templates and the methodology of minimal/distant supervised learning based on monolingual input avoid to have parallel text for training.
- Currently five languages are supported: German, English, Spanish, French, and Chinese.

Richard Ishida (W3C Internationalization Activity Lead) presented on “New Internationalization Developments at the World Wide Web Consortium”. The presentation focused on a specific aspect of W3C work: layout requirements, like justification in CSS for non-latin scripts, Japanese layout requirements, or layout for Korean or Chinese. The purpose of the talk was a call for action for the MultilingualWeb community:

Related links: Slides • IRC • Video

- The Open Web Platform needs much more information than is currently available to support the needs of world-wide cultures in user agents and standards. Local communities need to contribute that knowledge.
- There are people already working in various parts of the world on such information (groups working on Japanese, Korean, Chinese, Hindi, Latin layout requirements, and some interest from others), but they need additional support from experts and local stakeholders.
- Current pressing needs are for example getting consensus on Arabic and Asian approaches to justification, letter-spacing and similar features.
- we need experts to form groups to create requirements, but a much wider group of people can review what others are creating, map the differences between the requirements and current Open Web Platform technologies and propose changes. The public at large then needs to use

the emerging features.

Charles McCathie Nevile (Yandex), presented on "Multilingual Aspects of Schema.org". Schema.org provides a structured vocabulary that is interpreted by search engines including Bing, Google, Yahoo! and Yandex. The Schema.org vocabulary is already used within about 10% of Web content, and adoption is growing very fast. Significant remarks related to multilingual aspects:

Related links: Slides • IRC • Video

> Documentation of Schema.org is provided only in English. World wide adoption could be fostered via documentation in other languages and real multilingual examples.
> Yandex is in the process of making Schema.org more useful for the multilingual Web.
> Issues to resolve include the handling of names for structured information items: they include capitalization which is not common in many languages.

The developers session ended with a **Q&A** session. Several questions concerned infrastructure for cross-lingual information access and multilingual content creation. For example, could the Schema.org vocabulary help to foster cross-lingual information access on the Web? In Wikipedia, how can articles and even sub units (sections, paragraphs, sentences, ...) be synchronized across languages? Providing context information is important in both scenarios; for example, the appropriateness of a given translation may change over time.

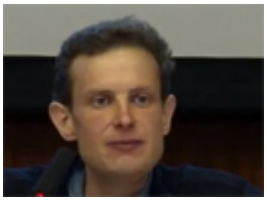
Related links: IRC

Creators session

This session was chaired by **Paul Buitelaar** of INSIGHT National Center for Data Analytics, National University of Ireland, Galway.







Fernando Servan (UN Food and Agriculture Organization FAO) presented "Bridging the Online Language Barriers with Machine Translation at the United Nations". The public website of FAO is

Related links: Slides • IRC • Video

using machine translation to provide content in Arabic, Chinese and Russian. The project started last year. Many sections are first created in English and then translated. Other significant remarks:

- Machine translation is used since translation services for other languages is sometimes hard to assure.
- Machine translation is implemented via a translation widget on the website, including a means to evaluate translation output.
- The project proved that machine translation can play a role for international, public organisations to provide a multilingual Web experience.

Marion Shaw, (SDL) held a presentation on "Multilingual Web Considerations for Multiple Devices and Channels – When should you think about Multilingual Content?". More and more users access the Web via multiple devices. Designing sites for these devices means: one has to consider multiple languages from the start. Other significant remarks:

Related links: Slides • IRC • Video

- Device manufacturers need to be aware that each new device feature can create new challenges for Web content localization.
- Content creators need to understand that the language of Web content is different than the language of many users who consume the content.
- The creation of structured content can help reuse during translation and lead to up to 30% cost savings compared to non-structured content.

Celia Rico, (Universidad Europea - Madrid), presented on "Post-editing Practices and the Multilingual Web: Sealing Gaps in Best Practices and Standards.". Post-editing of machine translation output has become a widespread activity in the translation/localization industry. There are still gaps to fill in terms of best practices and standards for post-editing. According to a study by Common Sense Advisory, the combination of machine translation and post-editing instead of human translation can lead to bigger translation volumes for the same price. Other significant remarks:

Related links: Slides • IRC • Video

- The study from Common Sense Advisory also states that the languages services market is growing at an annual rate of 5.13%.
- The standardized ITS 2.0 metadata can improve the speed and quality of post-editing.
- The usefulness of selected post-editing guidelines and ITS 2.0 has been analyzed in the EDI-TA project.

Gemma Miralles, SEGITTUR, gave a presentation on "Spain Tourism and Cultural Website". SEGITTUR is a company attached to the state secretary for tourism in Spain. It aims at promoting and implementing best practices and technical innovation in the domestic and international touristic markets. Its multilingual websites are an important means to achieve these goals. Other significant remarks:

Related links: Slides • IRC • Video

The three websites of SEGITTUR were launched at different times and differ with regards to multilingual content.
Multilingual content is created relying among others on translation companies, translation memories, style guides and glossaries.
Quality assurance in translation is an important step for the domain of tourism.
For this domain, it is important to translate for countries, not languages; e.g. one has to take into account how people search on the internet.

Alexander O'Connor (CNGL KDEG Trinity College Dublin) gave a presentation entitled “Marking Up Our Virtue: Multilingual Standards and Practices in the Digital Humanities”. Digital humanities allow for scientific methods that would not be possible without the digital representation. Important topics in digital humanities currently are text analysis, literary analysis, archives and repositories, data and text mining, and visualisation. Other significant remarks:

Related links: Slides • IRC • Video

Standards play an important role in digital humanities
Especially in the library domain standardized metadata helps to access digital artefacts, including multimedia objects.
There is a contextual challenge in a content and in the metadata, for example expression for entities (like names of cities) may change over time in history (see “Gdansk” versus “Danzig”).

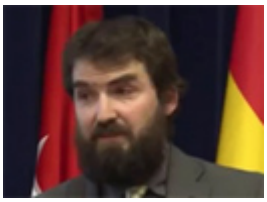
The **Q&A** session first focused on the topic of digital humanities. Using a computer to get research done faster is not enough; one has to address new topics that would not be possible outside the digital artifact. As another topic the language coverage of multilingual web sites was discussed. It depends very much on the domain of content in question and the content consumers targeted. Regional variants are relevant in some scenarios; the United Nations focus on standard language variants.

Related links: IRC

Localizers session

This session was chaired by **Phil Ritchie** from VistaTEC.



Jan Nelson (Senior Program Manager, Microsoft) presented “The Multilingual App Toolkit Version 3.0: How to create a Service Provider for any Translation Service Source as a key Extensibility Feature”. In this presentation, he discussed the XLIFF 2.0 standard and its role in Microsoft’s Multilingual App Toolkit Version 3.0. In particular, he emphasized the ways in which many translation sources can be easily integrated into app localization through the Toolkit and XLIFF 2.0. Other significant remarks:

Related links: Slides • IRC • Video

> Microsoft Windows currently covers 108 languages, with 10 languages providing 78% of coverage around the world. Adding languages increases coverage but is a time consuming process, which has to be simplified and to be made cost-effective.
> Language support is vital even in apparently monolingual markets. For example, in San Francisco, more than 45% of the population have non-English mother tongues.
> In a live demo, Jan Nelson demonstrated how to add translation sources to the Multilingual App Toolkit and how the translations can be leveraged across multiple versions of an application (e.g., Windows for PC, mobile, web). The tools make it easy to integrate terminology, MT, and human translation.

Joachim Schurig (Senior Technical Director, Lionbridge) gave a presentation entitled “The Difference Made by Standards-Oriented Processes”. He discussed the important role that XLIFF 1.2 plays in Lionbridge’s operations and the major benefits it can offer to a Language Service Provider (LSP). Other significant remarks:

Related links: Slides • IRC • Video

> Lionbridge currently uses a mix of standards, including XLIFF 1.2, TMX, ITS 2.0, and SRX. At present XLIFF 1.2 is used for approximately 95% of Lionbridge’s content production (approx. 2 billion words). Relying on standards has created significant cost reduction within Lionbridge.
> While XLIFF 1.2 is not perfect, it gave Lionbridge some benefits: elimination of character corruption issues, prevention of document corruption, reduction of back conversation failures, control over the editing environment and automation without unexpected failures. These features have allowed Lionbridge to automate processes in a new system and increase total translation volume per year.

Rob Zomerdijk (SDL) discussed “Content Relevancy Starts with Understanding your International Audience”. Discovering relevant information cross-lingually can be extremely difficult. Examining social media can help companies to understand the factors that matter to their international audiences. By focusing on existing conversations in different markets, companies can understand these markets much more

Related links: Slides • IRC

quickly than through traditional market research techniques. Other significant remarks:

- Many organizations need to take care about the ways in which they provide vast amount of content. However, understanding their audiences is very expensive.
- Conversations citizens have on social channels such as forums and blogs, and remarks they make on micro-blog services like Twitter create new data sets that can be used to understand your citizens.
- Translation quality must be optimized so that content is easier to find by and more relevant to the audience.

Tatiana Gornostay (Tilde) spoke on "Towards an Open Framework of E-services for Translation Business, Training, and Research by the Example of Terminology Services". She covered the online portal "Terminology as a Service (TaaS)" that provides cloud services for terminology work, including some less common languages (e.g., Latvian, Polish). TaaS supports searching for and identifying terms in a variety of source file formats and has many databases in the backend to support user needs. Other significant remarks:

Related links: Slides • IRC • Video

- Tatiana presented several use cases in translation and terminology work that address gaps with regards to standards and the multilingual Web.
- Open cooperation between various members of the Language Technology community (Business, Academic and freelancers) could lead to better terminology services.

David Filip (CNGL at University of Limerick), **David Lewis** (CNGL at Trinity College Dublin), and **Arle Lommel** (DFKI) presented on "Quality Models, Linked Data and XLIFF: Standardization Efforts for a Multilingual and Localized Web". The speakers presented a three-part update on current consensus building and integration in standardization activities, and a vision for future development. They described how recent developments in standardization around XLIFF, translation quality, and linked data are influencing each other. Other significant remarks:

Related links: Slides • IRC • Video

- XLIFF 2.0, currently in the final stages of candidacy as an OASIS standard, is a redesign of XLIFF 1.2 and improves various aspects. David Filip explained the new characteristics of XLIFF 2.0 and its roadmap for further development.
- One of the problems in the language industry today is that there is no agreement about methods of quality translation and that methods for assessing human and machine translation are totally disconnected. The QTLaunchPad project has worked on a standard definition of quality translation, which is fundamental to the new Multidimensional Quality Metrics (MQM) specification.
- Dave Lewis presented information about the basic principles of ITS 2.0 and current developments of representing ITS 2.0 information within linked data.

During the **Q&A** session questions were raised about whether it would be possible to move pass the current XML-based, bilingual paradigm for language processing. The consensus was that multilingualism will come in the future, but for now bilingualism will remain the default for some time, although there will be a shift to on-demand access to data. On additional problem is that the current multilingual resources tend to be academic in focus, but the LIDER project can do more to educate potential users about what can be accomplished with multilingual data. Finally, there was discussion about the move from file-based services to service-oriented architectures and how quality processes will have to adapt to this shift.

Related links: IRC

Machines session

This first part of this session on Wednesday afternoon was chaired by **Dan Tufis** of RACAI. The second part on Thursday morning was chaired by **Hans Uszkoreit** of DFKI.

Roberto Navigli (University of Rome) spoke on "Babelfying the Multilingual Web: State-of-the-Art Disambiguation and Entity Linking of Web Pages in 50 Languages". The presentation discussed the role of BabelNet, a wide-coverage multilingual semantic network in 50 different languages, in promoting semantic disambiguation and linking to named entities across language boundaries on the Web. The benefits would be considerable for multilingual communication since language technology could leverage Babelnet data to improve translation. Other significant remarks:

Related links: Slides • IRC • Video

> There is a lot of textual context in many different languages. BabelNet could be the basis to achieve high-performance multilingual text understanding.
> Babelnet 2.5 version is an integration of Wordnet, Wikipedia and other resources. Babelnet integrates all definitions for a given concept.
> The Babelfy tool provides word sense disambiguation and entity linking.

Victor Rodríguez Doncel (Universidad Politécnica de Madrid) gave a presentation entitled “Towards High-Quality, Industry-Ready Linguistic Linked Licensed Data”. The presentation focused on aspects of licensing of linguistic resources available for industry use. Different licensing models have different potentials and impact the possible business opportunities for multilingual linked licensed data. Work on linguistic linked licensed data is being carried out as part of the LIDER project. Other significant remarks:

Related links: Slides • IRC • Video

- In the so called Linguistic Linked Data Cloud (LLD), linguistic resources are being made available following linked data principles.
- Licensing of linked data is an important prerequisite for making business with data, without the fear of losing control.
- The topic of linguistic linked licensed data is one of the topics being discussed in the W3C’s Linked Data for Language Technology Community Group.

Seth Grimes (Alta Plana Corporation) spoke on “Sentiment, Opinion, and Emotion on the Multilingual Web”. The presentation provided an overview of the ways in which sentiment is conveyed on the web. Grimes challenged standard positive-negative scalar views of sentiment and discussed the complex relationship between sentiment and resulting action. There is no single method or representation of sentiment. Other significant remarks:

Related links: Slides • IRC • Video

- There are four varieties of data (machine, interactions and translations, profiles, and media) and two “super-types” (facts and feelings). Each one has its own characteristics, and the boundaries (particularly between facts and feelings) are not always clear.
- Increasingly sentiment is conveyed through non-textual means. Machine processing of non-textual input is currently improving.
- Multilingualism is increasing and adding to the complexity, especially because sentiment may not translate (e.g., use of “sick” as American slang for something good).

Asunción Gómez-Pérez (Universidad Politécnica de Madrid) spoke on “The LIDER Project”. LIDER works on the basis for creating a Linguistic Linked Data cloud, which can support content analytics tasks of unstructured multilingual cross-media content. In this way, LIDER will help to improve ease and efficiency of using linguistic linked data sources for content analytics processes. Other significant remarks:

Related links: Slides • IRC • Video

- LIDER aims to establish a new Linked Data ecosystem of free, interlinked, and semantically interoperable language resources and media resources.
- Various linguistic linked data formats are relevant for LIDER, e.g. Lemon for representing RDF-based lexicons, the W3C provenance ontology, or ODRL to represent licensing information, and NIF for corpora.
- LIDER works in the W3C Linked Data for Language Technology Community Group and collaborates with several other initiatives inside and outside of W3C.

Martin Brümmer (University of Leipzig) **Mariano Rico** (Universidad Politécnica de Madrid) and **Marco Fossati** (Fondazione Bruno Kessler) spoke on “DBpedia: Glue for All Wikipedias and a Use Case for Multilingualism”. This presentation focused on the central role of DBpedia in encoding knowledge derived from Wikipedia. The information provided via DBpedia can help to improve various language related technologies. Other significant remarks:

Related links: Slides • IRC • Video

> Martin Brümmer explained DBpedia’s goal of extracting knowledge from Wikipedia.
> Other groups, such as Unicode Consortium’s ULI, are using DBpedia to improve language related standards and technologies.
> Currently DBpedia has 14 languages chapters.
> The Italian and the Spanish DBpedia chapters presented several uses cases. Since the information is language- and culture-specific, an ongoing challenge is how to provide both universal knowledge and cultural specifics at the same time.

Jorge Gracia (Universidad Politécnica de Madrid) and **José Emilio Labra** (University of Oviedo) spoke on “Best Practises for Multilingual Linked Open Data: A Community Effort”. The presentation covered best practices for improving multilingual linked data sources, such as providing language tags. Work on such best practices is carried out in the W3C Best Practices for Multilingual Linked Data Community group (BPMLOD). Other significant remarks:

Related links: Slides • IRC • Video

> The creation of the BPMLOD group is a direct result of the 6th MLW workshop 2013, held in Rome. The BMPMLOD work is coordinated with the LD4LT group.
> BPMLOD is documenting patterns and best practices for the creation, linking, and use of multilingual linked data.

Hans Uszkoreit (DFKI) gave a talk about “Quality Machine Translation for the 21st Century”. The EU has made a major commitment to improve MT in order to support the demands of the European society. Currently there is an extremely uneven distribution technology support for various languages and MT quality is insufficient to support industry needs. The QTLaunchPad project, and the proposed followup, QT21, are seeking to address these needs by providing improvements to MT and other resources, working in cooperation with industry, to achieve the goals of a multilingual European society. Other significant remarks:

Related links: Slides • IRC • Video

> The European Single Digital Market is coming, but e-commerce in Europe faces many barriers.
> In addition to the 22 official languages, there are many other non-official languages that must be accounted for.
> There is a major gap in resources for different languages, and some of the most complex languages have the fewest resources.
> The QT21 project will focus on the following: substantially improved statistical and machine-learning based models for challenging languages and resources scenarios; improved

evaluation; continuous learning from mistakes; obtaining guidance from systematic analysis of quality barriers, informed by human translation; and learning and decoding efficient.

Related links: IRC

The discussion during the **Q&A** session was concerned with access to data. For example a lot of good work in the area of machine translation is lost because there is no central repository for data, resources, and results. Another topic was the relation between linked data and machine translation. Linked data researchers need to work with MT researchers to ensure that linked data can help to improve MT. In general, there needs to be more convergence between currently separate research themes. A number of languages are growing in importance, and the emergence of these languages reveals problems that have not previously been considered or solved.

Users session

This session was chaired by **Thierry Declerck** of DFKI.




Related links: Slides • IRC • Video

Pedro L. Díez Orzas (Linguaserve) started the Users session with a talk entitled “Multilingual Web: Affordable for SMEs and small Organizations?”. He presented a medium-term vision of factors that have an impact on the costs for multilingual web technologies, solutions, and services for a wide range of small and medium enterprises (SMEs) and public organizations. SMEs need to consider organizational factors when implementing these technologies. Other significant remarks:

The vast majority of LSPs (especially in Europe) are small and medium enterprises. Multilingual web technologies must support the needs of these companies in a scalable fashion.

One of the biggest shifts will be in moving from manual, “copy and paste” and email-based workflows to fully automated, on-demand systems.

Human factors (access to trained staff, knowledge, and resources) cannot be ignored. Standards can help in this aspect be reducing the reliance on proprietary solutions.

Related links: Slides • IRC • Video

Don Hollander (APTLD) gave a presentation entitled “Universal Access: Barrier or Excuse?”. The adoption of internationalized domain names (IDN) is needed in the whole software ecosystem, such as email clients, web browsers or mobile devices. Other significant remarks:

Don Hollander introduced APTLD, which administers top-level domains (TLDs) in the Asia-Pacific region.

The structure of TLDs has changed dramatically in recent years. TLDs may now have more than three characters and are no longer limited to latin characters.

User acceptance of TLDs is impacted by the fact that they do not work reliably in all contexts.

Dennis Tan (VeriSign, Inc.) talked about "Internationalized Domain Names: Challenges and Opportunities, 2014". The presentation provided detailed information about internationalized domain names (IDNs), with a focus on the growth in recent years and the challenges to adoption. He discussed efforts to break past bad user perception so that IDNs can grow over time. Other significant remarks:

Related links: Slides • IRC • Video

At present IDN penetration is relatively low: 98% of domain names are ASCII, with 1% of domain names using "extended" latin script and 1% using non-latin script. However, more than 50% of internet users speak non-latin script languages.

Many IDNs are implemented as redirected links that resolve to ASCII domains.

None of the world's most popular websites allow use of IDN email addresses, meaning that the public find them unusable for many purposes.

Awareness of IDNs is slowly increasing, particularly in Asia, but breaking the negative cycle of poor awareness and user experience will require a concerted effort.

Georg Rehm (DFKI) presented on "Digital Language Extinction as a Challenge for the Multilingual Web". He discussed the current state of support for language technology in Europe's languages and the impact that poor support has on European culture, heritage, and civil society. According to the META-NET study "Europe's Languages in the Digital Age", 21 of the 31 European languages assessed are in danger of "digital extinction", i.e., they will not be useable for widespread communication on the Web. Other significant remarks:

Related links: Slides • IRC • Video

The META-NET Strategic Research Agenda (SRA) provides a roadmap and concrete steps to address the threat of digital extinction.

Europe needs to focus on R&D that can support smaller, under-resourced, and complex languages.

Interdisciplinary collaboration and research is needed to overcome Europe's language barriers so that Europe can benefit from linguistic diversity and combat digital extinction.

In the **Q&A** session the role of IDNs was discussed, as well as how to facilitate use of multilingual web technology for smaller enterprises.

Related links: IRC •

Open space session

This session was chaired by **Arle Lommel** (DFKI). The discussion focused on the following issues:

Related links: Slides

Data formats. ITS 2.0 can be used for XML and HTML5. Several participants said that ITS 2.0 metadata needs to be available for JSON and other formats as well. To realize this goal, several issues need to be resolved, e.g. how to integrate ITS 2.0 information inline within JSON. The mapping of ITS to RDF may help to solve this issue.

IDNs and IRIs. At present there is little demand for companies to fix support for IDN/IRI since they need to see it as a problem that is worse than the cost of working around it. The group recommended that those who need these capabilities actually use them to break systems and then report the breakage to developers in an effort to raise awareness. The solution is likely to require action from many parties since all systems need to work together and even one weak link can make IDN/IRI useless to the end user. One major threat to the Web community is that if the major players do not solve the problem themselves, they run the risk of having problematic requirements forced upon them via legislation, so developers need to consider what will happen if they delay acting.

Future of the MLW Workshop series.. Considerable discussion focused on the future of the workshop series. After the end of the LIDER project there is no institutional support for the Workshops. Various options to continue were discussed, but the final consensus was that the workshop should remain free and keep the same format. The financial model will likely have to shift to more sponsorship, and it was suggested that “micro-sponsorship” options (e.g., 100€) be added so that those with budget can help support the program. The biggest costs will be in administration, but the program committee can help promote continuity at low costs. Additional suggestions on how to improve the workshop included: provide more guidance on how to use the online tools and consider providing an “etherpad” or similar document that can be collaboratively edited during the workshop; consider integrating a pecha-kucha session; consider better ways to manage questions (possibly using the online tools) and include ways to ask for clarification of points during the talks while deferring substantive questions to the end; look to increase the geopolitical diversity of speakers and include more women.

LIDER Roadmapping workshop

The LIDER roadmapping workshop consisted of seven parts: a keynote presentation from Seth Grimes, two panels on multilingual aspects of Wikipedia, a use case and requirements session, and three sessions on linked data and language technologies.

Related links: Roadmapping workshop page

Text Analytics, Applied. In his keynote presentation on Text Analytics Applied, Seth Grimes introduced approaches and usage scenarios of text analytics. Masses of unstructured content in many different forms (Web pages, emails, multimedia content, social media etc.) contains named entities, facts, relationships, sentiment information etc. Text analytics turns these into structured information for various usage scenarios. The presentation and the related report on text analytics provide information on application areas, the type of information actually being analysed in industry, or current language coverage. Important points for LIDER roadmapping were:

- Text analytics generates the structured information for bridging search, business intelligence and applications.
- Key real-life application areas are: business intelligence, life sciences, media & publishing, voice of the customer, marketing, or public administration and policy making.
- Users of text analytics tools need: adaptability to their content domain, customization (e.g. import of taxonomies), flexible input & output formats and processing mechanisms (API, offline, ...).
- Sentiment resolution is an important functionality for many usage scenarios.
- When deciding on solutions, users take capacity (volume, performance, latency) and cost into account.
- Multilingual text analytics is an area that so far has not seen a lot of activity in industry.

Using Wikipedia for multilingual web content analytics. In this panel session, Pau Giner, Amir Aharoni and Alolita Sharma provided details about the Wikipedia translation infrastructure. So far users only have indicators, but no explicit provenance information about what article is a direct translation from other languages. There is also no strict translation workflow. This is due to the nature of the Wikipedia community, including Wikipeida editors, translators and new users who do content creation via translation. Significant points of the session are summarized below.

- For its content translation tool, Wikipedia is looking into automatic translation tools, allowing the user to translate per pagraph and revise the result.
- Handling feedback from users for the tool development is a challenge since about 200 communities have to be taken into account.
- Wikipedia based machine translation tooling could help to quickly create domain specific MT.
- The multilingual infrastructure of Wikipedia could be the basis for new research topics like comparing content across cultures and in this way cultures themselves.

Growing Wikipedia editing with intelligent multi-language suggestion lists for article translation as well as other techniques and tools. This session started with Runa Bhattacharjee, Pau Giner and Santhosh Thottingal on a panel. It was highly interactive, which is reflected by the summary of key points below.

- Information on translation quality could help translators in Wikipedia. Such information is being defined in the QTLaunchPad project, see the presentation at the MultilingualWeb workshop from Arle Lommel.
- Various types of background information can help translators in several ways: providing translation suggestions, disambiguate terms, autocompletion etc.
- Data models for structured information in Wikipedia, e.g. Wikidata, do not rely on the linked

data, that is RDF technology stack. But conversions to and from linked data are possible
One challenge is the integration of resources like Wikidata or others that have been discussed at the MultilingualWeb workshop, like BabelNet or DBpedia.

Gathering of use case and requirements.

To further investigate and understand the use cases and requirements of LOD the participants were asked to participate in a post-it sessions. This sessions involved two sets of post-its, one related to use cases and one related to requirements. The participants provided approx. 40 different use cases and requirements. Based on the main theme being LOD most use cases and requirements reflected core LOD topics such as supporting MT through access to large (annotated) data sources.

However, interestingly larger trends around Linguistic data were identified and can be summarized as 'Access', 'Connectivity', 'Contribute' and 'Understand'.

'Access' relates to several comments indicating the need for more open APIs and data sources that can be used to support and augment existing data sets and applications. 'Connectivity', a core theme in LOD, was highlighted by the need for more RDF based data sources. 'Contribute' was indicated by comments seeking a stronger support and contribution by communities and community driven projects (e.g. Wikimedia and Wordnet). Finally, 'Understand' relates to needs identified in the are of sentiment and content analysis.

It can be concluded that most use cases and requirements related to providing access to more structured/annotated data sources and allowing simple connectivity via APIs and standards.

Initiatives and community groups related to linked data and language technologies. In this session key persons provided an overview of various groups. The groups presented are LIDER, FALCON, the LD4LT Community group, the OntoLex Community group, the Best Practices for Multilingual Linked Open Data (BPMLOD) Community Group and the Open Linguistics Working Group of the OKF. Key points of the discussion were:

One has to be careful in classifying resources. Some are language resources (e.g. lexica), others are general resources like Wikipedia / DBpedia / Wikidata. All of these can be relevant for linguistic processing, but they are different in nature.
An example of this aspect is Wikipedia. It is a knowledge based often used in natural language processing, but not a linguistic resource.
A (diagram of an) linguistic linked data cloud needs to clearly distinguish between different types of resources, since it is an entry point for potentially interested people or institutions.
The quality of resources is sometimes hard to evaluate. Looking at scientific publications can help, e.g. a resource mentioned frequently may be of interest.

Data and Metadata of Language Resources as Linked Data on the Web. The aim of this session was to discuss general approaches and concrete examples of how to represent language resource metadata and the resources themselves as linked data. Christian Chiarcos discussed ISO TC 37 originated standards like LMF or GrAF and RDF as a representation model. He when discussed various parts of the linguistic linked data cloud (corpora, lexical resources, term bases) and use cases and certain challenges for representing these as linked data.

Philipp Cimiano presented several use cases in the form of queries on linguistic resources, and

presented a proposal for a linked data based architecture to relize these use cases. Interoperability of metadata is a key challenge to make this vision happen. The LD4LT community group is working into this direction. One key metadata vocabulary is META-SHARE. Stelios Piperidis provided background on META-SHARE and its role in the planning of a new machine translation initiative QT21, see the presentation from Hans Uszkoreit at the MultilingualWeb workshop for details. Marta Villegas introduced the current state of mapping the META-SHARE XML Schema to RDF.

Roberto Navigli provided detailed technical aspects of BabelNet, complementing his general overview at the MultilingualWeb workshop. BabelNet relies on existing linked data vocabularies like Lemon, SKOS or LexInfo 2.0. RDF modeling issues arise when linking linguistic information to general knowledge e.g. in Wikipedia. Thierry Declerck showed how dialectal dictionairies can be represented as RDF and integrated with general linked data information. The encoding of textual information associated with entries and senses (e.g. examples) is a challenge.

Multilingual Corpus transformation into Linked Data. Here, the RDF based NIF format can help. Martin Brümmer introduced NIF. It aims to foster interoperability between natural language processing tools, language resources and annotations. In this way, NIF can serve as a pivot format for corpora. Roberto Navigli presented a concrete corpus that shows the application of NIF and ITS 2.0 information: an RDF version of the XML-based MASC corpus. Felix Sasaki demonstrated an application of NIF integrating both linguistic information and localization workflow metadata.

The above presentations clearly demonstrated the feasibility of NIF based corpus creation and processing. Laurette Pretorius closed the workshop discussing opportunities and challenges for representing information related to under-resourced languages of Southern Africa as linked data. The processes of publication, discovery, sharing and consuming of language resources could greatly benefit from linked data. To make this happen, proper tool infrastructure needs to be in place and best practices on how to work with linguistic linked data need to be made available.

Authors: Nieves Sande, Arle Lommel, Felix Sasaki. Report review: Richard Ishida. Scribes for the workshop sessions and for LIDER roadmapping: Guadalupe Aguado-de-Cea, Paul Buitelaar, Thierry Declerck, Dave Lewis, John McCrae, Kevin Koidl, Jorge Gracia, Eva Méndez, Roberto Navigli, Felix Sasaki. Photos in the collage at the top and various other photos, courtesy Volker Agüeras Gäng. Collage by Nieves Sande. Video recording and editing by Volker Agüeras Gäng.
