**LIDER: FP7 – 610782**

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

**Deliverable number**  
D4.9

**Deliverable title**  
Fifth Roadmapping Workshop Report

**Main Authors**  
Tiziano Flati, Jorge Gracia, Roberto Navigli, Felix Sasaki

<table>
<thead>
<tr>
<th>Grant number</th>
<th>Agreement number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>610782</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Project ref. no</th>
<th>FP7-610782</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project acronym</td>
<td>LIDER</td>
</tr>
<tr>
<td>Project full name</td>
<td>Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Starting date (dur.)</th>
<th>1/11/2013 (24 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ending date</td>
<td>31/10/2015</td>
</tr>
<tr>
<td>Project website</td>
<td><a href="http://www.lider-project.eu/">http://www.lider-project.eu/</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coordinator</th>
<th>Asunción Gómez-Pérez</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>Campus de Montegancedo sn. 28660 Boadilla del Monte, Madrid, Spain</td>
</tr>
</tbody>
</table>

<p>| Reply to              | <a href="mailto:asun@fi.upm.es">asun@fi.upm.es</a> |
| Phone                 | +34-91-336-7417 |
| Fax                   | +34-91-3524819 |</p>
<table>
<thead>
<tr>
<th>Document Identifier</th>
<th>D4.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Deliverable</td>
<td>LIDER EU-ICT-2013-610782</td>
</tr>
<tr>
<td>Version</td>
<td>1.2</td>
</tr>
<tr>
<td>Document due date</td>
<td>31 July 2015 (after extension)</td>
</tr>
<tr>
<td>Submitted</td>
<td>30 July 2015</td>
</tr>
<tr>
<td>Responsible</td>
<td>W3C/ERCIM</td>
</tr>
<tr>
<td>Reply to</td>
<td><a href="mailto:fsasaki@w3.org">fsasaki@w3.org</a></td>
</tr>
<tr>
<td>Document status</td>
<td>Final</td>
</tr>
<tr>
<td>Nature</td>
<td>O(Other)</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU(Public)</td>
</tr>
<tr>
<td>WP/Task responsible(s)</td>
<td>Felix Sasaki, DFKI / W3C Fellow</td>
</tr>
<tr>
<td>Contributors</td>
<td>Tiziano Flati (UNIROMA), Jorge Gracia (UPM), Roberto Navigli (UNIROMA), Felix Sasaki (DFKI / ERCIM)</td>
</tr>
<tr>
<td>Distribution List</td>
<td>Consortium Partners</td>
</tr>
<tr>
<td>Reviewers</td>
<td>Reviewed by the project consortium</td>
</tr>
<tr>
<td>Document Location</td>
<td><a href="http://lider-project.eu/?q=doc/deliverables">http://lider-project.eu/?q=doc/deliverables</a></td>
</tr>
</tbody>
</table>
Executive Summary

This document, the fifth roadmapping workshop report, summarizes the outcome of five roadmapping activities:

- On 17 March 2015, LIDER held two sessions at the SHARE-PSI 2.0 workshop in Timișoara. These sessions, a follow-up to LIDER participation in the SHARE-PSI 2.0 workshop in Lisbon (see LIDER D4.8 for details), provided valuable, additional feedback from the European public sector information community.
- On 29 April, 2015, LIDER organised the 8th MultilingualWeb workshop in Riga. The event was part of the Riga Summit on the Multilingual Digital Single Market. Like previous MultilingualWeb workshops, the event attracted a broad community of stakeholders around the multilingual Web.
- In the week of 15 June, LIDER organised the 1st Summer Datathon on Linguistic Linked Open Data (SD-LLOD-15). The event attracted a huge variety of participants from different areas in academia and industry.
- On 19 June, LIDER organised two sessions during the BDVA Summit 2015. The topics were “Multilingual Data Value chain” and “Standardisation”; the latter session was co-organised with non-LIDER contributions from W3C.
- On 6 July, LIDER organised its 5th roadmapping workshop in Rome. Different to previous workshops, the event focused on a specific community: cross-media linked data.

Like the LIDER deliverable D4.8, this report summarizes both roadmapping workshops and other events, which contribute to general LIDER roadmapping. All roadmapping activities held by LIDER are listed 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.
### Document Information

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

<table>
<thead>
<tr>
<th>Deliverable Number</th>
<th>D4.9</th>
<th>Title</th>
<th>Fifth Roadmapping Workshop Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workpackage Number</td>
<td>4</td>
<td>Title</td>
<td>Community building and dissemination</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Delivery</th>
<th>Contractual 31 July 2015 (after extension)</th>
<th>Actual 30 July 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>version 1.2</td>
<td>final ■</td>
</tr>
<tr>
<td>Nature</td>
<td>prototype □</td>
<td>report □ dissemination ■</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>public ■</td>
<td>consortium □</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authors (Partner)</th>
<th>Felix Sasaki (DFKI / W3C Fellow), Jorge Garcia (UPM), Tiziano Flati</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsible Author</td>
<td>Name Felix Sasaki E-mail <a href="mailto:fsasaki@w3.org">fsasaki@w3.org</a> Partner DFKI / W3C Fellow Phone +49-30-23895-1807</td>
</tr>
</tbody>
</table>

### Abstract (for dissemination)

This document, the fifth roadmapping workshop report, summarizes the outcome of three roadmapping activities: Two sessions at the SHARE-PSI 2.0 workshop in Timișoara. The 8th MultilingualWeb workshop in Riga, part of the Riga Summit on the Multilingual Digital Single Market. The 1st Summer Datathon on Linguistic Linked Open Data (SD-LLOD-15).

Two sessions during the BDVA Summit 2015, with the topics of “Multilingual Data Value chain” and "Standardisation"; the latter session was co-organised with non-LIDER contributions from W3C. The 5th LIDER roadmapping workshop in Rome, focusing on the specific community of cross-media linked data.
### Keywords
- LIDER
- roadmapping workshop
- report

<table>
<thead>
<tr>
<th>Version</th>
<th>Modification(s)</th>
<th>Date</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>First Draft</td>
<td>20/07/15</td>
<td>Felix Sasaki, DFKI / W3C Fellow</td>
</tr>
<tr>
<td>02</td>
<td>Review by InfAI</td>
<td>24/07/15</td>
<td>Bettina Klimek, InfAI, Asunción Gómez-Pérez, UPM</td>
</tr>
<tr>
<td>03</td>
<td>Final version</td>
<td>30/07/15</td>
<td>Felix Sasaki, DFKI / W3C Fellow</td>
</tr>
</tbody>
</table>
# Project Consortium Information

<table>
<thead>
<tr>
<th>Participants</th>
<th>Contact</th>
</tr>
</thead>
</table>
| Universidad Politécnica de Madrid | Asunción Gómez-Pérez  
   Email: [asun@fi.upm.es](mailto: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 Dublin, Ireland) | David Lewis  
   Email: [dave.lewis@cs.tcd.ie](mailto:dave.lewis@cs.tcd.ie) |
| Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany) | Felix Sasaki  
   Email: [felix.sasaki@dfki.de](mailto:felix.sasaki@dfki.de) |
| National University of Ireland, Galway (NUI Galway, Ireland) | Paul Buitelaar  
   Email: [paul.buitelaar@deri.org](mailto:paul.buitelaar@deri.org) |
| Institut für Angewandte Informatik EV (INFAI, Germany) | Sebastian Hellmann  
   Email: [hellmann@informatik.uni-leipzig.de](mailto:hellmann@informatik.uni-leipzig.de) |
| Universität Bielefeld (UNIBI, Germany) | Philipp Cimiano  
   Email: [cimiano@cit-ec.uni-bielefeld.de](mailto:cimiano@cit-ec.uni-bielefeld.de) |
| Universita degli Studi di Roma La Sapienza (UNIVERSITA DEGLI STU, Italy) | Roberto Navigli  
   Email: [navigli@di.uniroma1.it](mailto:navigli@di.uniroma1.it) |
| GEIE ERCIM (ERCIM, France) | Felix Sasaki  
   Email: [fsasaki@w3.org](mailto:fsasaki@w3.org) |
Table of Contents

1  INTRODUCTION.......................................................................................................................... 8

2  LIDER SESSION AT SHARE-PSI EVENT ................................................................................. 9
   2.1  INTRODUCTION.................................................................................................................. 9
   2.2  CONTRIBUTIONS .................................................................................................................. 9
   2.3  KEY POINTS OF THE SESSIONS ..................................................................................... 10

3  1ST SUMMER DATATHON ON LINGUISTIC LINKED DATA (SD-LLOD-15) ....................... 11
   3.1  INTRODUCTION................................................................................................................. 11
   3.2  CONTRIBUTIONS ............................................................................................................... 11
   3.3  KEY POINTS OF THE DATATHON .................................................................................. 20

4  LIDER CONTRIBUTIONS TO THE BDVA SUMMIT 2015 .................................................... 21
   4.1  INTRODUCTION................................................................................................................. 21
   4.2  CONTRIBUTIONS .............................................................................................................. 21
   4.3  KEY POINTS OF THE LIDER CONTRIBUTIONS ............................................................. 23

5  THE 5th LIDER ROADMAPPING WORKSHOP .................................................................... 24
   5.1  INTRODUCTION................................................................................................................. 24
   5.2  CONTRIBUTIONS .............................................................................................................. 24

6  CONCLUSION AND NEXT STEPS ......................................................................................... 31

APPENDIX I: MULTILINGUALWEB WORKSHOP REPORT ......................................................... 32
1 Introduction

This document, the fourth roadmapping workshop report, summarizes the outcome of five roadmapping activities:

- On 17 March 2015, LIDER held two sessions at the SHARE-PSI 2.0 workshop in Timișoara. These sessions, a follow-up to LIDER participation in the SHARE-PSI 2.0 workshop in Lisbon (see LIDER D4.8 for details), provided valuable, additional feedback from the European public sector information community.

- On 29 April, 2015, LIDER organised the 8th MultilingualWeb workshop in Riga. The event was part of the Riga Summit on the Multilingual Digital Single Market. Like previous MultilingualWeb workshops, the event attracted a broad community of stakeholders around the multilingual Web.

- In the week of 15 June, LIDER organised the 1st Summer Datathon on Linguistic Linked Open Data (SD-LLOD-15). The event attracted a huge variety of participants from different areas in academia and industry.

- On 19 June, LIDER organised two sessions during the BDVA Summit 2015. The topics were “Multilingual Data Value chain” and “Standardisation”; the latter session was co-organised with non-LIDER contributions from W3C.

- On 6 July, LIDER organised its 5th roadmapping workshop in Rome. Different to previous workshops, the event focused on a specific community: cross-media linked data.

Like the LIDER deliverable D4.8, this report summarizes both roadmapping workshops and other roadmapping activities, which contribute to general LIDER roadmapping.

All roadmapping activities held by LIDER are listed at https://www.w3.org/community/l4lt/wiki/Lider_roadmapping_activities

This page will be kept up to date during the duration of the project.
2 LIDER Session at SHARE-PSI Event

2.1 Introduction

SHARE-PSI 2.0 is a network for innovation in European public sector information (PSI). The main instrument of activity is a series of workshops to gather feedback from the PSI community and beyond. In December 2014, LIDER participants held two sessions at the workshop on encouraging open data usage by commercial developers (3-4 December, Lisbon). The outcome is documented in LIDER deliverable D4.8. The Share-PSI thematic network kindly invited LIDER to participate in the Timișoara SHARE-PSI workshop that took place in March 2015. The topic of this workshop was: Open Data Priorities and Engagement — Identifying data sets for publication. Like the workshop in Lisbon, the event consisted of parallel “un-conference style”, free discussion sessions which were held in parallel. For this reason, the below summary does not represent the flow of discussion or separate presentations. In total the LIDER sessions in Timișoara had about 10 participants with different type of background and interest.

2.2 Contributions

The two sessions organised by LIDER had the following foci. First, we discussed aspects of multilingual (linked) data and its role for identifying data sets for publication. Second, we went into more detail what technology is needed to work with multilingual linked data sets. LIDER provided the following background material to the participants:

- Some background slides to introduce LIDER and LLD.
- The LIDER D4.8 deliverable summarizing the multilingual sessions during the SHARE-PSI event in Lisbon, December 2014.

To trigger the discussions, five questions were presented to the participants.

1. What kind of products or services do you provide?
2. What kind of markets are you focused on primarily (financial, chemical, biomedical, …)?
3. Is multilingual data a challenge in your business? Do multilingual issues block your entry into markets in other countries? Any languages in particular?
4. Do you develop or buy language resources and/or tools to address the problem? Do you use linguistic open data sets? Do you see any problem with open data? Would you pay for linguistic data?
5. Do you think that a more standardized approach to language resources and/or tools will benefit your entry into other markets/countries? Do you know about or already use linked data and/or linguistic linked data?

The type of participants and their feedback is summarized as follows:

- PSI experts want to know learn about existing platforms, tools etc. for working with multilingual linked data.
- Linked data practitioners have been involved in linked data projects and some of them know already several aspects of linguistic linked data.
- Drivers of national data portals have the need to make the data multilingual.
- Providers of linked data tooling for public services are interested in automation in the realm of multilingual linked data, e.g. via machine translation.
• Developers of certain data sets, e.g. large taxonomies used for information interoperability in the public sector, have the need to make these data sets multilingual.

A general feedback from the participants was that a lot of public data is statistical data. Using language technology to translate the metadata for such data sets can improve access to the data.

### 2.3 Key Points of the Sessions

The outcome of the session was around best practices related to linguistic linked data. The participants agreed that LIDER should influence the W3C work on general data on the Web best practices. Related background material is the Data on the Web BP draft, ISSUE 5, and the background issue 142. An important source for this is the input from the W3C BPMLOD Community Group, and their various best practices, related to naming and dereferencing, textual information, linking, ontologies and vocabularies. Participants emphasised the importance of certain best practices, related to language resource vocabularies (e.g. for lexica or corpora), or URI design.

The beforehand mentioned developers of PSI data sets emphasised the need for both common and domain specific vocabularies.

A concrete next step for LIDER is to assure that multilingual data is a part of general data on the Web best practices. This will also help to continue and enlarge the community around LIDER beyond the end of the project.
3 1ST SUMMER DATATHON ON LINGUISTIC LINKED DATA (SD-LLOD-15)

3.1 Introduction

The 1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15) took place in Cercedilla (Madrid, Spain) from 15 to 19 June 2015, organized by the LIDER project. It was directed by Jorge Gracia from Universidad Politécnica de Madrid (Spain) and John McCrae from Bielefeld University (Germany). The main goal of the datathon was to offer persons from the industry and academia practical knowledge in the field of linked data applied to linguistics, with the final aim of allowing attendees to transform their own (or other's) linguistic data and publish it as linked data on the Web.

This datathon was the first organized on this topic worldwide. Around sixty five professionals (including 44 attendees, speakers and tutors) met in the event from all around the world. The datathon was an invaluable forum not only for learning but also for the exchange of experiences and ideas related to linguistic linked data. More information can be found at http://datathon.lider-project.eu/

3.2 Contributions

3.2.1 Program

The detailed program of the datathon can be seen in Figure 1. The datathon’s sessions were divided in four categories:

- **Invited talks.** Four selected invited speakers from outside the LIDER consortium were invited to give a talk (forty minutes followed by twenty minutes discussion) about a topic relevant to the datathon.
- **Seminars,** that were theoretical presentations (20 minutes + 10 for questions) given by LIDER members\(^1\) to show novel aspects and discuss selected topics.
- **Practical sessions** to introduce the basic foundations of each topic, methods, and technologies and where participants had the opportunity to do hands-on exercises, guided by the speakers and tutors. The required materials (software and data) were pre-installed in the computers of the datathon computer rooms, and distributed also on USB sticks in case the participants preferred to use their own laptops.
- **Datathon sessions,** in which participants, organised in groups and guided by tutors, planned and performed their own project on linguistic linked data.

---

\(^1\) With the exception of Gilles Sérasset, attendant to the datathon from Université Joseph Fourier, Grenoble (France) who was invited by the organisers to give a seminar about DBnary.
The invited speakers and their given talks were:

- **Rodolfo Maslias** (Head of the Terminology Coordination Unit, EU Parliament), “Institutional Terminology, Tools and Communication”. [Abstract](#) - [Slides](#)
- **Christian Chiarcos** (Goethe University), “Linked Open Dictionaries (LiODi) Lexical and phonological search in multilingual dictionaries”. [Abstract](#) - [Slides](#)
- **Marta Villegas** (Universitat Pompeu Fabra, Barcelona, Spain), “Publishing and Consuming Linked Data. (Lessons learnt when using LOD in an application)”. [Abstract](#) - [Slides](#)
- **Piek Vossen** (VU University Amsterdam, Netherlands), “The Global Wordnet Grid”. [Abstract](#) - [Slides](#)

All of these presentations captured a lot attention and motivated the debate both during the discussion part and later during the coffee breaks. The four speakers were invited to stay longer, so they had the opportunity of participating in the rest of activities of the datathon and could interact more with the participants.

This is the list of imparted seminars:

- **Asun Gómez-Pérez** (Universidad Politécnica de Madrid), “Maximising (Re)Usability of Linguistic Resources using Linked Data”. [Slides](#)
- **John McCrae** (Bielefeld University), “lemon: The Lexicon Model for Ontologies”. [Slides](#)
- **Felix Sasaki** (DFKI) “Roundtripping of NIF based Linguistic Linked Data with non linked data sources”. [Slides](#)
- **Jorge Gracia** (Universidad Politécnica de Madrid), “Apertium RDF: an experience in generating linguistic linked open data”. [Slides](#)
- **Philipp Cimiano** (Bielefeld University), “Linked Terminologies: applying linked data principles to terminologies”. [Slides](#)
- **Gilles Sérassett** (Université Joseph Fourier), “The DBnary eco-system, data and APIs”. [Slides](#)
And the practical sessions:

- Jorge Gracia (Universidad Politécnica de Madrid), “Introduction to Ontologies, RDF and LD”.
- Jorge Gracia and Daniel Vila-Suero (Universidad Politécnica de Madrid), “Multilingual LD generation and publishing”.
- John McCrae (Bielefeld University), “lemon”
- Ciro Baron and Bettina Klimek (University of Leipzig), “NIF”.
- Andrejs Ābele (Insight, NUIG), “RDF generation with D2RQ”.
- Tiziano Flati (Universitá di Roma “La Sapienza”), “BabelNet and BabelFy”.
- Mariano Rico (Universidad Politécnica de Madrid), “lemon-ade”.

Figure 2 shows a datathon practical session in progress.

![Figure 2: Work during a practical session.](image)

3.2.2 LIDER participants

From the LIDER consortium, six members volunteered to act as tutors: Gabi Vulcu (Insight, UNIG), Andrejs Ābele (Insight, UNIG), Víctor Rodríguez-Doncel (UPM), Tiziano Flati (UNIROMA1), Bettina Klimek (INFAI, University of Leipzig), Ciro Baron (INFAI, University of Leipzig). The tutors stayed in Cercedilla during the whole duration of the datathon. Every tutor had one or two datathon groups assigned to them and had the responsibility of monitoring their progress, assist them to clarify the group’s goals, assure that they followed a proper methodology, and help them with any possible technical issue (asking for assistance to other LIDER members if necessary).
There were also five LIDER speakers: Mariano Rico (UPM), Asunción Gómez-Pérez (UPM), Daniel Vila-Suero (UPM), Philip Cimiano (University of Bielefeld), Felix Sasaki (DFKI), Ciro Baron (InfAI) as well as one non-LIDER speaker: Gilles Sérasset (Université Joseph Fourier, Grenoble). They were in charge of giving some seminar talk or leading some practical session.

Most of the tutors acted also as speakers and gave either a seminar or some practical session. Also the datathon directors acted as speakers.

In addition to tutors and speakers, some other LIDER’s members with less commitment level acted as collaborators, to sporadically help tutors in their tasks and to assist in some logistic aspects. They were Thierry Declerck (DFKI), Guadalupe Aguado-de-Cea (UPM), and Elena Montiel (UPM).

Finally, another LIDER member, José Ángel Ramos, acted as datathon secretary, being in charge of all the administrative part.

### 3.2.3 Attendants

These are some basic statistics from the registered people:

- 44 participants (59 initial applicants)
- 24 different countries
- 34 cities
- 35% female 65% male
- 22% industry 78% academy

Figure 3 shows a world cloud with the different represented countries along with their relative importance (font size). A world map with the location of the different represented cities can be found in Figure 4 below.

![Word Cloud](image)

**Figure 3: Word cloud with the participant countries.**
There was no unique profile among the different registered participants: there were PhD students, developers from industry, research group leaders, university teachers, etc, coming from different areas: digital humanities, computer science, linguistics, etc. Also the participant’s previous experience was not uniform, ranging from little or no experience in linked data to people very experienced in semantic web technologies although willing to explore and contribute to the new LLOD paradigm.

We offered four travelling grants (up to 500€ each), intended for participants who could not cover their trip with other funds and giving preference to those coming from less-developed and/or distant countries.

The list of registered people can be found at http://datathon.lider-project.eu/#participants. In Figure 5 we show a group picture with all the datathon participants (attendants, tutors, speakers, and directors).
3.2.4 Developed projects

During one of the first sessions, the attendants were asked to organise themselves in working groups. Every group would have to select a leader/representative person and would have to decide on the particular topic and datasets to work with. To help in this task, the datathon directors suggested some group leaders and some possible topics, and split or merged unbalanced groups whenever it was necessary. Finally, nine working groups were organised with four to six members each one. This is a short overview of their final projects:

- **Group 1: GuanXi Networks.** The proposed system tries to overcome, with the use of linguistic linked data, the challenges of using linguistic resources in language learning and NLP (scattered data, lack of explicit meaning, etc.). They developed a multilingual LD network based on the integration of several resources (PDEV, Slovnyk, CEDICT, COW), using lemon and translation.owl as models. A linking method based on BabelFy was proposed and manually evaluated. Finally, a case study was described based on new Chinese words recently borrowed by English.

- **Group 2: Philological Lexicons.** This was about the conversion of the LIDDELL-SCOTT Greek dictionary from XML-TEI into RDF using lemon as model. Additionally, they converted a corpus in Latin (chartes bourguignonnes) to NIF.
and linked it to BabelNet. The motivation of the latter is to facilitate the discovery of concepts and topics in Latin texts, as well as harmonizing different tagging models (e.g., POS). Linking to BabelNet was challenging owing to orthography and cultural distance issues. In parallel, they also started converting a poetry repository as LLOD, with links to VIAF. Finally, they also planned the conversion of Lewis Short’s Latin dictionary and the Latin WikiQuotes.

- **Group 3: META-SHARE & LRE-Map.** The motivation of this work was LR metadata harmonization and reconciliation, in particular the Meta-Share and the LREMap datasets. ODRL was used to represent license data. They mapped the upper nodes of the ontologies into general categories and created a new ontology using Protégé. Then, the metadata was converted into RDF according to the new model. As future work, such metadata will be published as LLOD and integrated in LingHub.

- **Group 4: Terminology on Demand.** The motivating use case is a Spanish speaker who does not speak English and wants to extract some knowledge from twitter via SPARQL queries. The system performs term extraction and candidate translations from tweets based on IATE. Then, the original TBX is converted into RDF, and combined with the annotations generated by group 9. Both the IATE and twitter terminologies and annotations were in a triple store and could be queried.

- **Group 5: South African Languages.** This work was about converting multilingual agricultural data in South African languages as LLOD. The original data came from searchable PDFs, and were converted into CSV to allow their later processing. The extracted entities were searched in BabelNet and Falcons to establish external links. Eleven lexicons were created in RDF with added translations in English and some external links.

- **Group 6: Lemonification** of two language resources. Two resources were converted into RDF using lemon: a Swedish lexicon (saldo) and a dictionary from the Oxford University Press (OUP), each one having different license schemes. Links to DBnary and WordNet were explored. Some OUP dictionary examples were represented in NIF. An evaluation was carried out, based on a sense-sense DBnary-saldo based gold standard. The generated RDF was uploaded into the DBnary service and exposed in a SPARQL endpoint.

- **Group 7: K Dictionaries.** The project was about converting into RDF a multilingual dictionary (with Spanish as main node), initially in XML and with a privative license. A subset of the elements was left out for the conversion into RDF (some complex structures such as collocations, idioms, etc. that would need a more careful analysis). The models for the RDF representation were lemon, lexinfo and SKOS. The produced RDF was loaded in Fuseki and the results queried via SPARQL.

- **Group 8: Getty LOD Ontology localisation.** The goal was to convert a subset of the Getty LOD Ontology (Agent types: Artists) into RDF (using the translation.owl ontology) and localize it in as many languages as possible. They got translations
from several translation systems (including Google Translate, Baidu, Yandex, Bing) and combined them in order to rank the candidates. All the possible translations were included in order to allow for future user-based feedback. As next step, a disambiguation step should be included in order to avoid including noisy translations.

- Group 9: Semantic enrichment of Twitter. The goal was to provide enriched Twitter open data and to publish it as LOD. They used the Open American National Corpus of tweets, and used DBpedia and BabelFy for their semantic enrichment. They were converted into NIF and stored the resulting RDF in Fuseki. They worked in synergy with group 4.

After a voting among the participants, the GuanXi Networks project, by Group 1, was selected as the best datathon project and therefore declared as winner of the “best datathon result” award (600€, evenly split). Figure 6 shows a picture of the group members after receiving the prize.

![Figure 6: Winners of the "best datathon result" award with the datathon organizers after receiving the prize.](image)

### 3.2.5 Social aspects

One of the objectives of LIDER (and therefore of the datathon) has been community creation. To favour this, we promoted interaction among participants both through the social activities and the work in groups.

To that end we organised a “lightning talk session” on the first day, in which each attendant had to present their background and motivation in one slide in one minute, as a way to know each other better and to help with the group formation. Also in the first day we had an informal “icebreaking session” with a variety of social games.

In the second day we had an excursion in the surrounding area with a professional guide, in which we visited the surrounding woods, a Roman road, and a water reservoir. On Thursday, we visited the historical city of Segovia, having dinner in a prestigious local restaurant.
The Internet social networks also helped us to disseminate our event. For instance, the datathon activities and social aspects were disseminated in Twitter by using the #sdllod15 (and optionally #LiderEU) hashtags. See https://twitter.com/hashtag/sdllod15?src=hash

The event motivated also some blog posts by non-LIDER members, such as

- [http://www.maslias.eu/2015/06/one-big-cloud-all-terminology-all.html](http://www.maslias.eu/2015/06/one-big-cloud-all-terminology-all.html) (by Rodolfo Maslias, EU Parliament),
- [http://kaiko.getalp.org/about-dbnary/21-languages-are-now-available/](http://kaiko.getalp.org/about-dbnary/21-languages-are-now-available/) (by Gilles Séraasset, Université Joseph Fourier),

### 3.2.6 Participant’s opinions

According to the post-event survey\(^2\), most of the participants evaluated the event very positively, acknowledging the opportunities they had for learning and doing networking in an inspiring environment. In a three-degree scale, the organisation of the datathon was considered “very good” by the 95% of participants and “reasonable” by 5%. Nobody rated it negatively (“poor”). The feedback and assistance get by tutors was considered “very good” by the majority (84%) and “reasonable” by the rest (16%). Again, nobody rated it negatively (“poor”).

The majority of participants (90%) considered that the focus of the datathon was neither too academic nor too industry oriented, but “just right”. An 84% of participants considered that the atmosphere of the datathon was conductive to learning and a 100% considered it conductive to networking.

Regarding the type of sessions, most of the participants considered all the sessions enjoyable and beneficial to them. However we detected that they missed more time for purely practical activities. For instance, despite the duration of invited talks and seminars was considered “just right” (79% and 89% respectively), the duration of the practical Sessions and datathon sessions were considered “too short” (78% and 79% respectively). This is an improvement point to be considered in future editions of the datathon.

In terms of learning per topic, this is a summary table with the participants’ opinions on how much they did learn on each topic:

\(^2\) The satisfaction questionnaire was distributed online during the last day of the datathon and it was answered by 19 participants (43% out of the total).
Finally, the *social activities* were also very well rated. All the participants (100%) who answered the questionnaire considered that they benefited “a lot” from the social activities.

### 3.3 Key Points of the Datathon

In the following, we summarise the main outcomes and benefits of the datathon:

- Increase of awareness of Linguistic Linked Data.
- Community creation: the datathon attracted many people interested in linguistic linked data, putting them in connection both among them and with other experts in the field (invited speakers, tutors, etc.). The participants were encouraged to:
  - Join the relevant W3C and OKF community groups, specially [W3C Linked Data for Language Technologies](http://www.w3.org/2001/sw/wiki/LinkedData4LT) community group and [OKF Open Linguistics](http://openlinguistics.org) working group,
  - Keep the contact with the other participants (by email, LinkedIn, etc.),
  - Finish the work started in groups during the datathon and submit it to relevant workshops/conferences.
- Dissemination of guidelines and best practices for linguistic linked data. All the guidelines, reference cards, etc. generated in the context of the LIDER project and W3C community groups were disseminated among the participants. Also the Linghub aggregator of linguistic metadata was introduced to the participants.
- Through the projects developed by the participants, several benefits were obtained:
  - Increase of the (future) amount of linguistic linked data on the Web,
  - Identification of a number of potential applications and use cases for linguistic linked data technologies,
  - (Partial) development of such linked data based applications.

---

3 [http://www.lider-project.eu/guidelines](http://www.lider-project.eu/guidelines)
4 Deliverables D2.1.1, D2.1.2
5 [http://linghub.lider-project.eu/](http://linghub.lider-project.eu/)
6 The RDF generated in the projects will be eventually published as LLOD on the Web, but most of it needed some more elaboration at the time the datathon was finished. This has to be done by the own participants, with the assistance of LIDER members if required.
4 LIDER Contributions to the BDVA Summit 2015

4.1 Introduction

The Big Data Value Association (BDVA) is the industry-led counterpart to the European Commission for the implementation of the Big Data Value Public Private Partnership. BDVA fosters research, development and innovation around the topic of big data. The concrete goals of BDVA are documented in the BDVA Strategic Research and Innovation Agenda\(^7\). The SRIA is updated once a year to reflect new priorities around big data. In the current version of the SRIA, the topics of language and standardisation are mentioned as important in several areas. But at the moment, the topics are not described detailed to a level that may lead to concrete actions, e.g. in terms of research and innovation actions dedicated to multilingualism or standardisation, in terms of success indicators (e.g. “language coverage” as a key performance indicator for upcoming big data technologies) etc.

For LIDER, both topics are of crucial importance. Hence, during the first BDVA Summit, held in Madrid 17-19 June, LIDER organised two sessions: about multilingual data value chains and about standardisation\(^8\). The standardisation session was co-organised with the W3C data activity lead Phil Archer, to assure broad interest from the data community.

4.2 Contributions

4.2.1 Session on Multilingual Data Value Chains

About the participants

The session had 22 participants with a variety of interests, like:

• Aspects of integration, interoperability and standardisation;
• Multilingual industry tools for the data economy and multilingual service infrastructures;
• The role and interrelation of specific technologies like machine translation, machine learning and Linguistic Linked Data; and
• Application scenarios like multilingual and broadcast news production or information extraction from medical texts.

Industry aspects of Multilingual Data Value Chains

During the session several sectors were identified that can benefit from multilingual data value chains, like eCommerce, tourism and news / media. In general, SMEs need to internationalize their business. The challenges are both technical but also differences in regulation. There is a need for easy to use solutions that take non language requirements like security into account.

Technical challenges

Several technical research and innovation challenges were identified. The role of machine translation with regards to big data technologies should be explored. This encompasses the task of using machine translation to make linked data sources multilingual. Here, often only small items need to be translated and / or are available as

\(^7\) See http://bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria_v1_0_final.pdf for version 1.0 of the SRIA.

\(^8\) See the landing pages for the sessions at http://lider-project.eu/?q=multilingual-data-value-chains and http://www.w3.org/2015/06/bdva-summit_standardisation/.
training data for machine translation systems. Another challenge is how data mining, currently being language agnostic, should be enhanced with language technologies. The level of resource (e.g. training corpora) availability differs enormously between languages. The linguistic linked data (LLD) cloud is an important element to bootstrap the cycle between data creation, usage in machine translation training, and application of machine translation for improving the level of multilingualism in linked data.

Challenges across technologies for Multilingual Data Value Chains
A general prerequisite to foster research and application development for the multilingual data value chain is the standardisation of language resources. Standardisation should be driven by open source tooling to make adoption of standards easier. Several language technology solutions exist. But in practice a small group of solutions is preferred because it is easy to use. A one-stop shop with easy access to all language technology tooling could help to showcase value of multilingual technologies. This should include different providers for the same services including information about quality and rating. Efforts around the Connecting Europe Facility (CEF) may be a way to provide language technologies not only to public administration, but long term also to the European public at large. In order to feed multilingual topics into the big data SRIA, one needs to have concrete business cases that require multilingual technologies. The BDVA effort on support for SMEs could help with that.

4.2.2 Session on Standardisation

About the participants
The session on standardisation had 13 participants. An important success of the session was that it brought quite heterogeneous people together:

- Technology users and providers from various large and small sized industries;
- Representatives of standardisation bodies like W3C and ISO; and
- Researchers contributing to standardisation in the (multilingual) data area.
- Researchers and technologies in need of certain solutions implementing standards.

Liaisons and existing work
It was pointed out that there are already efforts in the area of big data standardisation, like the ISO/IEC JTC1 study group on big data. The report provided by the study group surveys existing technologies related to big data, identifies standards gaps, and proposes standardization priorities.

Suggestions around standardisation and big data
Standardisation itself should be funded. One should not expect that after technology development standardisation would just happen magically. Standards bodies follow innovation to reduce workload / help innovators. Public funding should be available for tooling implementing standards. Technology users need guidelines on how to use standardised aspects of big data technologies. On the other hand, big data developers and researchers need guidelines on how to use standards that are in place for solving particular problems. Examples of such guidelines are available, around the topic of creation and usage of multilingual datasets. One should not standardise processes but capture terminology used related to given and new processes. It is difficult to standardise processes themselves because innovation leads to new processes. However, capturing terminology related to given or new processes may be beneficial. For example, if people use the term “enrichment” within the content lifecycle with the same meaning, then it is easier to align different processes across organizations.
It must be made easy to find standards & technologies. This goal should be supported by efforts in dissemination about standards. The BDVA should list some of the most relevant Big Data standards in its SRIA, with the promise of pushing these. The focus should shift towards applying standards, with reduced emphasis on developing new ones.

To help the adoption of standards, one could put standards compliance into procurement. But SMEs need support for engaging in standardisation. They face a different situation compared to big companies. For SMEs, standard compliance is hard.

Like in the session on multilingual data value chains, use cases that need standardisation should be driven by SMEs, as the major economics and innovation driver. They can also provide business cases for the validation of standardisation activities.

**Technical areas that can benefit from standardisation**

In the content value chain, there are more and more resources like DBpedia, Babelnet, used in text analytics services that point to big data, aggregating a lot of knowledge sources. The problem is that these services are offered via many different APIs. So here is a need for standardised content analytics APIs.

Currently there are many different APIs for machine learning and predictive analytics. Again here is a need for API standardisation.

Many semantic standards are hard to understand and hard to process for analytics experts. Easy to use formats like JSON-LD can help to access such data. A common task may be conversion of existing data into such formats.

One should provide consensual terminologies in RDF to allow data exchange on European level. Again one should provide this in easy to process formats (like JSON, CSV, etc.) and automate conversions. In general one should take multilingual dimension in vocabularies into account.

Work is needed on interoperability between XML and linked data formats. This is of specific importance for certain sectors, e.g. digital publishing. Here we need interoperability between content formats like Pub (HTML5), XML and linked data.

A general requirement on technical new standards is: to assure that no existing infrastructures need to be changed.

Data curation and data quality need to be taken into account as standardisation areas. Big data more and more is being used to extract information about decisions / conclusions. For that we need data with assured quality. At least you need to know what is the quality and if it is compatible to other data sets. Therefore, one needs to make data sets comparable in terms of quality. We need frameworks for defining data quality; not sure if existing frameworks are prepared for big data. We may need a big data quality model, defining quality in given a context and for given purposes. This would allow for computing quality, in some cases at least, within and across application domains.

Finally, a question was raised on standards related/applied to Web data anonymisation. Data with personal information is needed for a lot of tasks related to big data, e.g. in the case of analytics of health related data.

### 4.3 Key Points of the LIDER Contributions

A key outcome of the standardisation session was that the BDVA decided to start a group to discuss standard related topics on a regular basis. The session on multilingual data value chains made clear that the role of multilingualism needs to be strengthened in the big data SRIA. LIDER will continue to push for this goal and to be a bridge between big data and the language technology related communities.
The 5th LIDER Roadmapping Workshop

5.1 Introduction

This report gives a summary of the 5th LIDER roadmapping workshop, which took place on 6th July 2015 in Rome, at the department of Computer Science of La Sapienza University of Rome. The topic of the workshop was “cross-media linked data” in the relation with the linguistic linked data. For more information, take a look the workshop program.

The main objectives were:

1. to put in contact several industrial stakeholders and public authorities interested in using linked data for multimedia content analytics;
2. to let them present and discuss their own technologies, problems, encountered obstacles, obstacles to be removed, visions, etc. about linked data for multimedia;
3. to introduce the LIDER project and the most recent version of the Linguistic Linked Data Cloud and to explain several prototypes to ease the exploitation of Linked Data.

In what follows, a summary of each talk will be presented.

5.2 Contributions

5.2.1 Welcome and Introduction

Roberto Navigli (Sapienza University of Rome) opened the workshop. After a short introduction of the EU LIDER project, the goals of this workshop have been outlined. The main focus was set on the use and needs of cross-media linked data for the business and industry sector and their connection with the linguistic linked data.

5.2.2 Gabriele Ciasullo (AgID): Open Data in the PA

Gabriele introduced the current panorama of Agenzia per l'Italia Digitale (AgID), the main public body which brings the open data paradigm into the national authorities and which fosters the policies for the enhancement of national public IT assets (Open Data by default). He presented the regulatory aspects which govern the re-use of public information and related tools for the concrete implementation of the processes for opening data by public administrations. He also described clauses which will allow the access and reuse of data in the definition of specifications or patterns of contracts for the collection and management of public data. A presentation of the evaluation of management performance followed. Gabriele also stressed the art. 68 of the CAD: public data, available data (licensed data), accessible data (technology which enables the automatic treatment of data), free data, open data. These are the tools that the public body uses for planning and checking its strategies: national agenda, guidelines, annual report.

Gabriele also introduced the next “National agenda 2015”: availability of public data according to the principles of open data, participation to work groups, new key datasets, starting processes for improving data quality, dynamic (Web-based) integration and update, national open data portal reengineering.

The “Annual report 2014” included the definition of certain dimensions of the OD: quality, availability and use/reuse of open data as well as indicators for each dimension.
According to the “Guidelines 2014”, a conceptual reference model for data and metadata should have been designed, the data production and release should have been modeled, and it was included also the creation of standards and ontologies for data and metadata, licensing, provenance, cost and publication of open data.

Gabriele afterwards described the PSI 2013/37/EU directive on public data, for commercial and non-commercial purposes and presents the obligation of member states to make documents reusable (art. 52 and 68 of the CAD). Future work was finally outlined.

Silvia Mazzini (Regesta): LOD projects in libraries archives and museums.

Silvia gave a general presentation of the company: Regesta, a company founded in 1996, started working with linked data at the Chamber of Deputies and kept working on historical archives.

Focus was put on the two linked open data projects made at Regesta, within the project “Libraries, Archives and Museums”, with a focus on the methodologies used for the publication, on the ontologies developed and the results obtained.

Silvia presented their catalogue software for archival heritage.

In 2010 they started working on semantic technologies to foster interoperability. Silvia presented the results obtained by the company she and Diego Camarda work in: they have achieved a good publishing quality, according to T. Berners-Lee classification, also as witnessed by the Grand Prize they have just won in Sydney for the LOD challenge last July. “We are currently collaborating with the Hebrew documenting center in Milan. At the core of our business we put LD technologies, because of the several advantages they bring.”

One of the latest projects deals with the reconciliation of the victim names during World War II. Names turn out to be ambiguous and data is extremely heterogeneous, bringing so many different additional data (e.g., pictures, documents of the historical site, books, postcards, etc.) and having so many different formats. The same person is associated with many different records (e.g., Primo Levi), since the information about that single person is replicated across the different offices which committed errors during the transfer into digital archives.

They have also worked on the creation of an ontology on the Shoah: data has been modeled by means of a specialized ontology and then transformed into triples, creating a persistent IRI for each resource, person, persecution, event, etc. Thanks to this process, everybody which has to deal with Primo Levi is able to use the IRI unique to that person. The LD also simplifies the data cleaning process because non-conformant data naturally emerges. “Shoah Victims’ Names” is now in the LOD cloud and the portal is already public. It provides also “sameAs” and “seeAlso” links, also towards datasets external to the considered domain. The interface shows internal as well as external data, something which is innovative.

Claudia Nicolai (FAO): Multilingual repositories at FAO: Opportunities for cross-media data access

The Food and Agriculture Organization of the United Nations (FAO) is one of the main publishers in the area of food and agriculture. The FAO Document Repository hosts more than 12000 publications (PDF or HTML) and most of them are available in the six UN languages (English, French, Spanish, Chinese, Russian, Hebrew), providing a significant corpus of multilingual data linked as metadata records in a database (with more than 7 million downloads per month). Similar FAO repositories exist or are in development for photographs, presentations and other multimedia data. A lot of multimedia data is available, with over 5 million users each month. “We produce more than 450 articles in English each year, 150 in French and Spanish, 30 in Hebrew, Russian and Chinese”.

FP7-610782

D4.9 Fifth Roadmapping Workshop Report
Although this information is available on the Web, unfortunately it is not linked in a way that users could make a cross-platform use of them. Metadata is created by means of keywords from a thesaurus (AGROVOC) and the document cataloging is unfortunately performed manually by data librarians, only on titles and abstracts. When the title is short and does not summarize the content, cataloging is performed at a lower level by going into the article’s text. This talk was aimed at describing the efforts being done to improve the retrieval of the information and exploring the case for enriched user experience with cross-media data access.

Claudia presented the in-house photo catalogues: metadata include keywords from a controlled vocabulary. Social medial channels from other multimedia content: youtube (videos), flickr (photos), slideshare (presentations). Metadata include hashtags (not-controlled vocabulary) and the real challenge is to put together all the pictures into a centralised repository.

Claudia pointed out the need for migrating to a digital asset management (DAM) software and for an identification (or development) of a taxonomy. Cross-media data access: documents and media. The current taxonomy comes with a number of issues, though. For example, 2013 has been the international year of the ‘quinoa’; a lot of documents have been produced about the ‘Quinoa’, but no central, unified interface is available yet. AGROVOC has not been able to solve the problems for two main reasons: 1) it is not flexible (e.g., ‘desertificazione’ and ‘desertizzazione’ are different terms, but they were conflated as synonyms) – and fixing this is very hard, 2) AGROVOC turns out to be too rich and too generic (e.g., the ‘Animal disease’ domain is very rich, “we have a huge database of animal autopsies decorated with a lot of text but very few labels”).

Future work: review of hashtags, normalisation, identification of keywords actually used, identification of existing taxonomies for describing multimedia content and collaboration with partners.

5.2.3 Andrea Volpini ([InsideOut10](#)): WordLift for the digital publishing and the construction of open dataset.

The [WordLift project](#) (the application for organizing and enriching the content on the Web) and the research activity, co-financed by the European project [MICO](#) (Media in Context) which is instead focused on cross-media analysis, have in common the goal to foster not only research and innovation but also technology at the service of bloggers and publishing companies. Technology is going toward integrated artificial intelligence (voice-enabled assistants, Hound project, Jibo, amazon echo). Furthermore, Associated Press has started to produce a huge amount of news automatically with AI techniques (sport, finance reports, etc.). “Text generation algorithms can interpret your data and turn it into meaningful, personalised content”.

Content published on the Web can be optimised by means of tools and methodologies which bring a semantic enrichment, through open linked data and the analysis of multimedia contents. WordLift is a semantic editor implemented as a WordPress plugin, also used together with editorial offices. It is intended for journalists as well as bloggers. It assists writers during their editorial process by suggesting context and enriching the content with images, links and widgets. Wordpress has been chosen for two reasons: on the one hand WordLift needs an underlying CMS to handle and store the pages associated with the recognised entities in the text, and on the other hand Wordpress turned out to be characterised by a very high reach.
Not only WordLift provides links, internal to the reader, and adds structured metadata but it also builds an open LOD dataset (entities, annotations, content) with the goal to create a structured organisation of the content, according to WHAT, WHO, WHERE and WHEN, also suggesting images and accessing LD domain-specific bases. The plugin also creates navigation systems to navigate through the content. The property dct:references is used to link text to concepts but also schema.org is used for basic knowledge (e.g., date of birth/death, etc). The WordLift system is under development on the Greenpeace magazine (used for fundraising) and offers content analysis (by relying on Apache Stanbol), content discovery (through Apache Solr) and linked data publishing (via Apache Marmotta). The current version relies on the AirPedia Named Entity Recognition and Disambiguation tool (ex-WikiMachine) but, according to the developer’s experience, it is hard to embed a general-purpose algorithm into arbitrary domain or context. The true added value is the possibility to have a custom vocabulary.

The MICO project is also considering the cross-media analysis: after the transformation into RDF, metadata models are defined to characterise the multimedia information and a battery of media extractors have been developed (e.g., animal detection, video quality, temporal segmentation, automatic speech recognition, face detection, audio tampering). The contribution of the project is represented by the introduction of SPARQL-MM (SPARQL MultiMedia) with the goal to bring multimedia features into SPARQL, also with spatiotemporal filters. SPARQL-MM will be available as a module of Apache Marmotta within October 2015.

Andrea: we rely on the AirPedia Named Entity Recognition and Disambiguation tool (ex-WikiMachine). Our experience tells that it is hard to embed a general-purpose algorithm into any domain and any context. The true added value is the possibility to have a custom vocabulary.

Diego: is there a particular reason why you choose to develop WordLift as a WordPress plugin?

Andrea: we chose WordPress because it has high reach. In any case we need a CMS to handle and store the pages associated with the recognized entities.

Gabriele: does the tool allow to verify the correctness of the multimedia query?

Andrea: No, it is out of our scope. SPARQL-MM will be available as a module of Apache Marmotta within next October.

5.2.4 Alexandros Chortaras (National Technical University of Athens - NTUA): Advancing the semantics of descriptive metadata in fashion collections

Good and rich descriptive metadata is of key importance for providing advanced search services over heterogeneous fashion collections. In Europeana Fashion (see http://www.europeanafashion.eu) a data semantification workflow engine has been developed in order to selectively apply ontology-based text analysis techniques on the textual descriptions, coming from more than 700.000 digital fashion metadata records aggregated by over 30 different European museums and archives, in order to identify references to specific semantic resources (e.g. AAT thesaurus terms and DBPedia location entities), and then enriching the analysed metadata by including links to the relevant detected concepts. Part of the data is represented by museum data (with rich metadata but expressed with other ontologies) and photographs (often, with no metadata).

The project, composed by 22 partners from 12 European countries, was aimed at creating a specialised access point for fashion content, developing case studies for the
exploitation of fashion material, building consensus and raising awareness on best practises.
The main features of the project included added fashion-related elements (e.g. color, material, designer), and use of specific relationships (e.g., between a dress and an exhibition). At the core of the project there is a multilingual thesaurus in 11 languages, for fashion and fashion-related concepts, including materials, techniques, object types, events and colors. The repository is developed as an aid for data entry, as a knowledge base (for internal and external users) and as a search enhancement tool.

Description of the Europeana portal: the data submitted by the providers have metadata fields defined in the EDM-fp schema (a SKOS-based taxonomy). Sometimes values are expressed as free text, in which case only keywords-based search is possible.

Semantically characterise free text descriptions: difficulties include i) multilingual data and ii) no training data. The domain is very strict, so ambiguity is not such an important problem.

Data linking includes linking external resources such as DBpedia.

Possible extensions: introducing axioms in the taxonomy, definitions of common terminology concepts and useful complex concepts. Semantic query answering to users asking for complex terms (Italian haute couture costume, flapper dress, extravagant dress, etc.).

5.2.5 Diego Camarda (Regesta): Lodview & LodLive: esplorare la linked data cloud

A first step toward the reuse of linked data resources is represented by their discoverability and navigability (value of the interoperability among linked data repositories). The W3C publication specifications allow to develop software that is able to explore and visualise the whole LOD cloud. LodView and LodLive have exactly this objective. The linked data approach allows to describe heterogeneous objects which is possible for extracting and displaying together with their metadata. The two open-source projects are designed to serve exactly this type of tools. As a result of querying LodLive about “La Sapienza”, the interface interacts by displaying a circle, surrounded by a series of additional circles (internal links) and roses (external links) which ease the data navigability from the user’s point of view. Guido finally presented the Lodview tool as a possible alternative to previous RDF data visualization tools, in contrast to old-style tools (e.g., DBpedia’s linked data front-end).

Alessio Bosca and Matteo Casu (CELI): Cross-media Linked Data and Soundscape Generation

The SoDA project (http://sodaproject.eu/) (Sound object Design Accelerator) proposes the design and implementation of a sound synthesis engine in real-time for sound design, interfaced with sound libraries through semantic analysis. The idea of the project is to provide sound designers with a simple, yet powerful, tool to accelerate their production process. In particular, the project provides a flexible software environment for soundscape generation and offers an integration of cross-media material on the basis of captions associated with pictures, videos, etc. Based on semantic information, it provides both an annotation schema and an annotated library of sound files that operates in relation to a generative system that delivers the final audio content. The annotation process can be performed both on a conceptual level (the type of sound; e.g., forest, cat, rain, etc.) and a technical level (frequency of a barking rather than of a machinegun). Matteo also queried the SoDA system with “forest rain with frogs” and played the integrated audio file obtained in this manner.

In the ongoing project “Librare” (http://www.librare.org/) Semantic and Internet-Of-Things technologies will be used in order to enhance small school libraries (little schools,
universities, up to cross-reading initiatives, etc.) and support didactic activities (e.g. social reading) by cataloguing and tracing physical and digital books. Matteo explained that Librare can be seen as an aggregator of existing systems such as “Librinlinea App Librare”, “App Librando”, “Cbook”, “FirstLife”, etc., and it also comes with the opportunity to map users to their geo-positions. It is a project which depends on the impact on the user network (i.e., how many users can be involved in Librare). Matteo finally presented “The Crunched Book”, a digital humanities Web application to read and explore literary works. The collected data will be published on the smart data platform of "Regione Piemonte".

Elena Candia (Camera dei Deputati – Chamber of Deputies): Publishing and consuming parliamentary Linked Open Data

The Italian Camera dei Deputati (Chamber of Deputies) has been adopting the linked open data paradigm since the 2010, to meet needs of interoperability, publishing the i) historical portal (storia.camera.it) - based on a LOD repository made of data and digital documents coming from several systems concerning more than 150 years of parliamentary history – and ii) the platform dati.camera.it for publishing all the open data obtained. On this line, the Chamber is publishing as LOD not only information about the current parliamentary activity but also information contained in the intranet, with particular focus on the management of internal documentation, with benefits in terms of increased internal efficiency, reduction of costs and improvement of integration in information sharing. The problem that has been internally encountered is that, despite LOD enabled users to add connections between these datasets (e.g., Gazzetta Ufficiale), the use of the LOD is now purely internal to prepare dossiers and preparatory documents for those who need more specific information on some topic (documentation center). The main problem that has been necessary to face was that of the visualisation of the data of the Chamber of Deputies: despite the conversion into linked data, the information in the website often needs to be presented using different formats. Future work: internal (for interoperability) and external (for transparency) reuse of LOD. “We are preparing legislatives thematically (department of the environment, department of work, department of social affairs, etc.) in which people used to access to registers which were little and proprietary”.

5.2.6 Antonio Basili (BabelMind): BabelMind: the Italian startup for centralized access to multilingual linguistic linked data

BabelMind is a new-generation company, founded in Rome in 2014 from the idea of a dynamic group of computer scientists, researchers and professionals whose goal is making human knowledge accessible in simple, effective and innovative ways. Going beyond the current landscape, BabelMind works on cutting edge multilingual semantic technologies. At the core lies the creation of a huge semantic network which stems from the research on BabelNet and goes well beyond: text analysis, text categorisation, entity recognition, sentiment analysis. BabelMind seamlessly integrates, extends and manipulates encyclopedic and dictionary knowledge from heterogeneous sources, enriched with audio and visual content. BabelMind provides a unified multilingual repository of knowledge for solving issues in many areas such as computer-assisted translation, localisation, multilingual semantic processing of texts, cross-lingual information retrieval.

In the context of the Gutenberg project, about 50 thousand texts in different languages will be disambiguated in order to extract the knowledge contained therein and represent this knowledge in the form of structured knowledge (e.g., graphs), also according to chapters, paragraphs, etc. The crucial aspect of the disambiguation process is the linking with BabelNet, which in turn enables multilinguality, and makes it possible to
enrich content with images, definitions, sounds and further information.

Round Table and Key Points of the Workshop

Roberto Navigli closed the workshop with a round table in which all the participants were invited to discuss actively. These are summarized as key points below:

Linguistic linked open data:

- How many were aware of the Linguistic LOD before this event? Only 3 people, Andrea, Alessio and Matteo. This shows that the workshop has reached people who did not work on the linguistic part of the linked cloud, an important aspect for evaluating the impact of this workshop.

- Elena: “a serious problem we encountered is that the images and all the multimedia content we have, regarding politicians, official events, and so on, are completely decoupled from the parliamentary acts. This type of data is associated with events, people, artworks: it should be very easy to associate concepts with these. It is critical to treat texts and integrate them as linked data or link them to videos, for example”. It is clear that the current linked data ecosystem is divided into different sections which hardly communicate: an integration of the experience of scientists in different areas is thus necessary for bringing the LOD cloud a step forward and achieving a true, linked cross-media environment.

- Alessio: raised the problem of closed linked data. “The BBC, for example, has been using linked data only for their selves, fundamentally (just as the Chamber of Deputies)”. Roberto: “Well, I think the future of linked data is strictly interconnected with licensing. After all, each piece of linked data has its own license which might be different from node to node. It will be up to agents to explore data according to their license”.

- Alessio: the Open Annotation Model is about to become a standard by the W3C and this is probably going to go in conflict with NIF. There is a clear need for establishing clear, well-defined standards which are able to cope with the inherent problems concerning the publication, exploitation and processing of linked data.

- A general feeling is that there is a rivalry between paradigms: for example who comes from the Semantic Web and LD is more interested in the conceptual level and those who come from computational linguistics are more interested in the words and their meanings. Andrea: “at the linguistic level, either I do have a very specific need (to disambiguate short fragments of text, e.g., tweets) or I am not going to use the disambiguation tool at all”.

A clear outcome of the workshop has been the need to develop tools which enhance and foster the integration of cross-media content with textual information, both via automatic (e.g., software) and manual (e.g., guidelines for the annotation) approaches.
6 Conclusion and Next Steps

The following table summarizes the outcome of LIDER community building activities in the 1st half of 2015. The 6th LIDER roadmapping workshop was held 13 July 2015 in Munich. Its outcomes will be summarized in LIDER deliverable D4.10.

<table>
<thead>
<tr>
<th>Event</th>
<th>Target communities and people reached</th>
<th>Outcomes</th>
<th>Next steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARE-PSI EVENT <strong>EVENT</strong></td>
<td>Open data; 10 key stakeholders from various groups (data providers and users)</td>
<td>More detailed feedback on potential benefits using LLD and needs</td>
<td>Potentially SHARE-PSI workshop contribution at SHARE-PSI workshop November 2015, to continue engagement after the end of LIDER project</td>
</tr>
<tr>
<td>8th MultilingualWeb workshop</td>
<td>Web community at large; around 100 participants</td>
<td>Raised broad awareness about LIDER outcomes</td>
<td>Continue linguistic linked data engagement with the Web community</td>
</tr>
<tr>
<td>Summer Datathon on Linguistic Linked Data</td>
<td>Linguistics, Digital Humanities, NLP, Semantic Web; both from Academia and Industry</td>
<td>Community creation. Raised awareness and adoption of LIDER outcomes (guidelines, best practises, reference cards,...). Development of applications on top of the LLOD cloud.</td>
<td>Support the finalisation of the developed LLOD-based projects in the next weeks/months. Find a way to support a second edition of the datathon (high interest was shown by the community)</td>
</tr>
<tr>
<td>BDVA Summit sessions</td>
<td>Big data community at large; 300+ participants</td>
<td>Installation of a group around standardisation in BDVA</td>
<td>Assure that multilingualism and standardisation are taken into account in Big Data SRIA update</td>
</tr>
<tr>
<td>5th LIDER roadmapping workshop</td>
<td>Companies &amp; Academics; around 15 people</td>
<td>Analysis of cross-media usage within Italian public bodies and projects in private companies</td>
<td>Potential collaboration with the Chambers of Deputies for improving the publishing and production of linked data and multimedia content</td>
</tr>
</tbody>
</table>
Appendix I: MultilingualWeb Workshop Report


Here we have added a PDF version of the online report. It is recommended to read the online version since it contains additional links to slides, videos etc.
W3C Workshop Report: Data, content and services for the Multilingual Web
29 April 2015, Riga

Today, the World Wide Web is fundamental to communication in all parts of life. As the share of English web pages decreases and that of other languages increases, it is vitally important to foster multilingualism for 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 29 April 2015 the W3C ran the eighth workshop in the series. The theme of the workshop was Data, content and services for the Multilingual Web. The workshop in Riga was co-organized by Tilde and co-located with the Riga Summit 2015 on the Multilingual Digital Single Market. The opening session was organized as a joined session with CEF – Towards a Connected Multilingual Europe. The following presenters gave a brief welcome address: Dace Melbarde (Minister of Culture of the Republic of Riga), Márta Nagy-Rothengass (Head of Unit, Data Value Chain, European Commission, DG Connect), Jānis Kārkliņš (Internet Governance Forum) and Richard Ishida (World Wide Web Consortium, Internationalization Activity Lead).

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 to meet the challenges of the multilingual Web. The key objective was to provide the opportunity for networking across a wide range of communities.

Participants were able to switch between the workshop and the parallel CEF event. Together the two events had more than 200 registered participants and featured one day of talks and discussions. A specific focus of the MultilingualWeb workshop was on data, content and services on and for the multilingual Web, provided via standardized technologies.

All presentations were recorded, and are available on the Web using the MultilingualWeb YouTube channel. We also provided scribing of presentations and discussions. The related session notes are linked from the presentations.

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

After a short summary of highlights, this document provides a description of each talk accompanied by a selection of key messages. In addition there are links to session notes, video recordings of the presentations, and slides.

During the workshop we recognized the contributions of Jörg Schütz. Jörg was a long-term contributor to the MultilingualWeb community. He was supposed to present at the Riga workshop and passed away unexpectedly a few weeks before.
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 MultilingualWeb, Grant Agreement No. 250500, and by the LT-Web project, Grant Agreement No. 287815.

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.

During the initial joint session held with the parallel CEF event, Dace Melbarde (Minister of Culture of the Republic of Riga), stressed the importance of user friendly, multilingual digital services especially for minority languages, while Mártá Nagy-Rothengass (Head of Unit, Data Value Chain, European Commission, DG Connect) talked about the imminent deployment of digital services infrastructure. Jānis Kārkliņš (Internet Governance Forum) discussed how to foster multilingualism on the internet e.g. via broad adoption of IDN (Internationalized Domain Names), and Richard Ishida (W3C, Internationalization Activity Lead) argued that standards create interoperable diversity, and should be world-wide in scope, not limited to Europe.

The keynote presentation from Paige Williams stressed how multilingual technology is about changing the lives of people around the world. Paige gave a broad overview of technologies that are crucial for a truly multilingual Web. The workshop continued with the Developers and Creators session and Han-Teng Liao. He emphasized the need for best practices and standards that put the user into the center. Roberto Navigli and Rodolfo Maslias presented a multilingual metasearch service, which allows users to access Web information in multiple languages.

Fernando Serván presented work undertaken by the Food and Agriculture Organisation (FAO) in the area of multilingual publishing. There is a need to implement culture specific requirements for high quality layout. Juliane Stiller closed the session with a presentation on the evaluation of multilingual features in Europeana.

The Localizers session was opened by Leonid Glazychev. He discussed methodologies for standardizing quality assessment, a topic that was taken up later by Arle Lommel, Jan Nelson presented on a concrete tool implementing localization standards: the multilingual app toolkit. David Filip made a presentation with contributions from Loic Dufresne on localization technologies being deployed within Intel: the so-called I18n/L10 service bus.

The Machines session started with a joint presentation from Asunción Gómez-Pérez and Philipp Cimiano. They presented intermediate outcomes of the LIDER project, among other items focusing on a roadmap for the use of linguistic linked data in content analytics applications. The presentation from Andrzej Zydron and Dave Lewis provided the current state in the FALCON project. The machines session closed with a presentation from Ilan Kernerman on multilingual glossaries.

The following session of lightning talks started with Felix Sasaki. He presented the FREME project, which is developing interfaces to several language and data technologies for multilingual and semantic enrichment of digital content. Felix also presented on behalf of Phil Ritchie the Ocelot application. Ocelot is going to implement several enrichment functionalities of FREME. The last lightning talk was given by Ben Koeleman and introduced the topic of swarm translation.

The final, Users session started with Delyth Prys, who talked about best practices for supporting minority languages, using the example of Welsh. Thibault Grouas introduced JocondeLab, a multilingual web site providing access to cultural heritage artifacts via linked data technologies. The Users session finished with a presentation from Dennis Tan. With the example of internationalized domain names (IDN), Dennis emphasized the need to take an
end-to-end perspective into account, involving every part of multilingual content creation, distribution and consumption. This view fit very well as a closing message to the workshop and supported the holistic and broad perspective of the MultilingualWeb community.

**Keynote presentation**

Related links: [IRC](#) • [Video](#)

The keynote was given by **Paige Williams** (Microsoft). She spoke about People-First: Multilingualism in a Single Digital World. Paige pointed out that in today's world of borderless communication using a growing amount of heterogeneous technologies, one has to rethink the classic approach to localization. Technology has to be made available in the preferred language of a global audience. Language technologies play an important role for achieving this goal. By improving global communication, language technologies can also contribute to growth in a global economy. Other important remarks:

- Standards simplify the process of working with multiple languages. Microsoft is building bridges between language, culture and technologies through language related standards and the Microsoft local language program.
- Microsoft provides the global readiness approach for employees and external developers. This should help them to build adequate customer experience for all markets and languages.
- Paige encouraged the audience to embrace diversity and to try to understand what local and global really mean.

Related links: [IRC](#)

During the Q&A session, questions were raised about Microsoft's plans to provide speech technologies for small languages and for regions with little or no internet connectivity. Paige replied that currently technology support for smaller languages is low, and that Microsoft sees this is an important challenge to be addressed in the future.

**Developers and Creators session**

The developers and creators session was chaired by **Tatjana Gornostaja**, Tilde.

Related links: [Slides](#) • [IRC](#) • [Video](#)

**Han-Teng Liao** (Oxford Internet Institute) gave a presentation entitled A call to implement a common translation and country-language selector repository. In his view, there are several gaps when it comes to expressing country and language choices in a consistent and usable fashion. These gaps are related to preferences around machine translated content, and language or country preferences. Europe with its great cultural and linguistic diversity should foster the development of related best practices. Other significant remarks:

- The Unicode [Common Locale Data Registry](#) (CLDR) provides a lot of machine readable information needed to address the challenges being discussed.
- European institutions should push for the harmonization of auto-translation mechanisms and language and country selectors. This concerns both the public and the private sector.
- Public institutions that could provide good examples are immigration departments or visa and border control agencies.

Related links: [Slides](#) • [IRC](#) • [Video](#)

**Roberto Navigli** (Sapienza University) and **Rodolfo Maslias** (European Parliament) gave a presentation entitled Metasearch in a Big Multilingual Terminological Space. In the European
Union every day a vast amount of multilingual data is being produced. Roberto and Rodolfo presented work on a multilingual metasearch engine that could help to tackle this challenge. The metasearch engine integrates several large-scale, multilingual data sources. Other significant remarks:

- Domain specific data is key to make the search results more accurate. The metasearch engine uses data from domains like labor market or migration.
- The metasearch engine could play an important economic role, e.g. by helping with cross border and cross lingual job search.
- Multilingual term bases are a crucial element for implementing these and other search scenarios.

Related links: Slides • IRC • Video

Fernando Serván (FAO) gave a presentation entitled Moving from a Multilingual Publications Repository to eBook Collections in the United Nations. FAO is publishing high volumes of multilingual content every year. The publications are made available using the standardized ePub format. FAO wants to take high quality layout requirements for a wide range of cultures and languages into account. Other significant remarks:

- At FAO, currently the creation of ePub documents is not fully automated. Each ePub document needs to be corrected manually before publication.
- Due to this situation, FAO needs standards and best practices on how to implement high quality layout requirements in the area of digital publishing.
- To reach a global audience, especially requirements from non-latin scripts and typography traditions need to be taken into account.

Related links: Slides • IRC • Video

Juliane Stiller (Humboldt University) gave a presentation entitled Evaluating Multilingual Features in Europeana: Deriving Best Practices for Digital Cultural Heritage. She reported on recent improvements in Europeana, the European digital library. A key feature of Europeana is cross-lingual querying of information about cultural artifacts. Juliane reported e.g. about improvements made in the area of query translation. Other significant remarks:

- A standardized multilingual vocabulary for identifying entities like persons, locations, events etc. is a basis for implementing cross-lingual queries in Europeana.
- Enriching metadata about cultural artifacts with links to entity information can improve search results tremendously.
- A non-technical challenge is the re-use of suitable multilingual resources, including licensing aspects.

Related links: IRC • Video

The developers and creators session ended with a Q&A session. Related to Hang-Teng's presentation, it was pointed out that European public administration, specifically the Publications Office of the European Union, has some high quality metadata that can help to tackle the challenges described. Also related to Hang-Teng's presentation, it was pointed out that W3C provides some best practices on language and country selection. These best practices could be improved with additional information from various layout traditions.

Data used by the multilingual metasearch engine (cf. the presentation from Roberto and Rodolfo) was discussed as well. The metasearch engine relies both on data sources from the public sector, like IATE, and from industry. For successful metasearch, one needs three categories of database: public data, academic data and data from industry.
Localizers session

This session was chaired by Fernando Serván of the Food and Agriculture Organization of the UN (FAO).

Related links: Slides • IRC • Video

Leonid Glazychev (Logrus Intl.) gave a presentation entitled Standardizing Quality Assessment for the Multilingual Web. Leonid discussed the need for a standard way to assess translation quality. He introduced a proposal for a standard: ASTM WK46397. It aims at simplifying quality assessment. Other significant remarks:

- The proposal covers several aspects like quality assessment via crowdsourcing, the general assessment process, and quality metrics.
- Leonid presented an application of the proposed standard for the review of a well-known public web portal.
- He emphasized the need for standardized metrics to assess translation quality, a topic covered later by Arle Lommel.

Related links: Slides • IRC • Video

Jan Anders Nelson (Microsoft) gave a presentation entitled XLIFF 2.0 and Microsoft’s Multilingual App Toolkit. Jan gave an overview of Microsoft's approach towards localization. Several years ago support of standards was not an important element of the company strategy. This has changed, as can be seen by Microsoft’s effort on XLIFF 2.0. Jan demonstrated the role of XLIFF 2.0 in the multilingual app toolkit for localizing applications, targeting several platforms like Windows, iOS and Android. Other significant remarks:

- The use of XLIFF 2.0 enables developers to re-use large percentages of their code.
- Developers also benefit from translation services that are provided via the toolkit.
- During the presentation he showed how cross-platform projects are supported by Microsoft’s use of XLIFF 2.0
- The support includes both machine translation services as well as engagement with language services providers.

Related links: Slides • IRC • Video

David Filip (University of Limerick), gave a presentation entitled Developing Standards-Based Localization Service Bus at Intel. David presented also material from Loic Dufresne de Virel. The Intel localization group has started a partnership with the Irish ADAPT centre. The aim of the collaboration is to design within Intel a data model and architecture for Internationalization and Localization: the I18n/L10n service bus. Other significant remarks:

- The I18n/L10n service bus benefits from supporting several recent standards in the realm of multilingual content production.
- The modularity of standards is a prerequisite to match abstract requirements of business processes.
- Relevant standards encompass CMIS 1.1, ITS 2.0, XLIFF 2.0 and the upcoming XLIFF 2.1.
Machines session

This session was chaired by Feiyu Xu from DFKI.

Related links: Slides • IRC • Video

Philipp Cimiano (Bielefeld University) and Asunción Gómez-Pérez gave a presentation entitled LIDER: Building Free, Interlinked and Interoperable Language Resources. They presented an overview of the intermediate outcomes of LIDER. The aim of LIDER is to build the basis for a linguistic linked data (LLD) cloud. LIDER produces guidelines and best practices for the creation of LLD, tooling like LingHub that demonstrates the application of linked metadata for exploring linguistic resources, and a community around LLD and use cases and requirements for content analytics tasks of unstructured multilingual cross-media content. Other significant remarks:

- The LIDER roadmap for the use of LLD Data for content analytics tasks is being developed with valuable feedback from various research and industry communities. These are active in several W3C community groups: BPMLOD, OntoLex and LD4LT.
- The LIDER reference architecture describes how to develop LLD applications based on existing and new multilingual resources and their deployment in natural language processing services.
- LIDER as a project plays a crucial role in bridging communities like language technology and (big) data technologies.

Related links: Slides • IRC • Video

Andrzej Zydron (XTM International) and Dave Lewis gave a presentation entitled FALCON: Building the Localization Web. They introduced the current state of the FALCON project. FALCON is deploying the LIDER principles of linguistic linked data in the translation and localization community. The project cooperates closely with LIDER to provide feedback from this industry on LLD use cases and requirements. FALCON has developed an LLD enabled online translation workflow that combines technologies from the realms of translation, translation management, computer-aided translation, and terminology management. Other significant remarks:

- The FALCON tool chain has been enhanced with automatic text extraction, machine translation and publicly available language resources.
- These components are integrated into one workflow through a web services architecture. It leverages open standards and linked tabular data formats.
- FALCON demonstrates how statistical machine translation training can be integrated with manual translation correction by human translators.

Related links: Slides • IRC • Video

Ilan Kernerman (K Dictionaries) gave a presentation entitled Semi-Automatic Generation of Multilingual Glossaries. K Dictionaries provides multilingual dictionaries as a basis for glossary generation. The semi-automatic generation process can be enhanced via linguistic linked data technologies. Standardized formats for representing lexica as linked data like LEMON can help to improve the process. Other significant remarks:

- Publishing houses, including providers of lexica, over several years have invested in XML technologies. Hence, the creation of linked data sources often involves a custom conversion from XML to RDF.
- Sub steps of the conversion include e.g. processing of word lists and continuous re-processing of the results.
- The multilingual glossaries help to improve access to web content across languages.
During the Q&A the level of disambiguation in multilingual glossaries was discussed. This is an area of active development and improvement for multilingual glossaries. Here resources like BabelNet can be of great help. The availability of BabelNet and multilingual glossaries as linguistic linked data eases the task of resource integration.

Participants were asking for actual use cases of linguistic linked data. As examples, references were made to the case of BabelNet and IATE, see the presentation from Roberto Navigli and Rodolfo Maslias, and the use of linked metadata by the Publications Office of the European Union. The session chair Fei Xu described how in German national Big Data projects the relation between big data technologies and natural language processing is being explored. So-called sar-graphs help to build a bridge between linguistic and world knowledge.

**Lightning Talks session**

This session was chaired by David Filip from University of Limerick.

**Arle Lommel** (DFKI) gave a presentation entitled Designing Purpose-Specific Quality Metrics for the Web. He discussed the Multidimensional Quality Metrics (MQM) and their application to web content. The ITS 2.0 Localization Quality Issue data category is an application of a subset of the quality types provided by MQM. Certain tools like Ocelot allow to parameterize the actual set of metrics being used for a given quality assessment task.

Related links: Slides • IRC • Video

**Felix Sasaki** (DFKI/W3C Fellow) gave a presentation entitled FREME project: Language and Data Processing as First-Class Citizens on the Web. FREME is a European project that which aims bring various types of language and data technologies to the market. This is done via the design of interfaces (software interfaces for programmers and graphical user interfaces) for implementing multilingual and semantic enrichment processes for digital content. The interface design is driven by four selected business cases and may inform future standardization related to language and data services on the Web.

Related links: Slides • IRC • Video

On behalf of Phil Ritchie (VistaTEC) Felix Sasaki gave a presentation entitled Ocelot: An Agile XLIFF editor. The presentation provided an overview of the Ocelot tool, a flexible XLIFF editor that is an integral component of Okapi framework, an open source set of components and applications for localization purposes. Ocelot's flexibility can be seen in the forehand mentioned adaption to MQM, as well as its usage within the FREME project. In FREME currently Ocelot is being adapted to allow translators for deploying multilingual semantic enrichment during localization.

Related links: Slides • IRC

**Ben Koeleman** (YAYANGO) gave a presentation entitled Swarm Translation. Ben shared his experience in a special translation project: in a huge community of volunteers, over 25,000 volunteers collaboratively translated books into their native language. Ben called this approach swarm translation. An important aspect of swarm translation is that this is not a sequential process. Many passages are translated simultaneously, including parallel discussions on the translations themselves. Ben shared lessons learned concerning the technological and organizational set up of swarm translation. Formulating these as best practices may help to guide future, similar endeavors.

Related links: IRC • Video
In the Q&A session, one challenge of swarm translation was discussed: how to balance demand and translation service offers. This is difficult in a community project that consists of volunteers. Related to the FREME project, priorities of the interfaces that are being developed were discussed. To make linked data accessible to people who are not linked data experts, one needs interfaces that hide complexity of data sets and data access, and the ability to work with non linked data, structured formats like CSV. The usefulness of tools like Ocelot or translate5 was questioned, since they don't provide what translators need in daily life, that is, CAT (Computer-assisted translation) tooling functionality. This topic has to be separated from the design of interfaces to certain data and language technologies provided by FREME. The vision behind these interfaces is to ease technology integration across platforms.

**Users session**

This session was chaired by Olaf-Michael Stevanof of ASLING and JIAMCATT.

Related links: Slides • IRC • Video

**Delyth Prys** (Bangor University) gave a presentation entitled Best Practices for Sharing Language Technology Resources in Minority Language Environments. She presented a new Welsh National Language Technologies Portal launched by the Language Technologies Department of Bangor University. The portal provides technology components for developing language technology applications for the Welsh language. Example components are spelling/grammar checkers, part of speech taggers and a machine translation system. Other significant remarks:

- Having technology components available is not enough. To achieve uptake by technology users one needs documentation, tutorials and examples of how to use the technology.
- Another crucial aspect for minority languages is the costs of language technology components. The portal for Welsh makes the components available for free, to foster commercial application development.
- The portal for Welsh can be seen as a success story of how to avoid digital extinction - a term that was created to explain the situation of smaller languages, see the META-NET White Paper Series.

Related links: Slides • IRC • Video

**Thibault Grouas** (Ministry of Culture and Communication, France) gave a presentation entitled Building a Multilingual Website with No Translation Resources. Thibault introduced JocondeLab, a multilingual web site that provides access to cultural artifacts in France. A key feature of JocondeLab is that it provides information integration of various types of multilingual resources, using linked data technologies. Among these resources are the french DBpedia. Interlinking this and other resources, JocondeLab provides access in 14 languages. Other significant remarks:

- The approach of creating multilingual user experience via publicly available cross-lingual linked data sources can be seen as a best practice of low budget multilingual content creation.
- The multilingual experience was created without actual translation (by humans or machines), but purely based on cross-lingual linking. The only translation / localization task concerned the multilingual user interface.
- JocondeLab won the French Data Intelligence Award. For re-use, the data has been made available under various licenses.

Related links: Slides • IRC • Video

**Dennis Tan** (VeriSign, Inc.) gave a presentation entitled Towards an End-to-End Multilingual Web. Dennis presented on the current state of internationalized domain names (IDNs). IDNs
allow users around the world to experience web addresses in their own languages and scripts. There is a huge demand for IDNs, since most users are living in countries and regions that use non-latin scripts. Due to the interconnected nature of the Web itself, the tooling around IDNs needs to be made available in an end-to-end manner. Other significant remarks:

- Sometimes certain issues like homographs are regarded as a challenge hindering adoption of IDNs. Dennis pointed out that there is mature technology to avoid the danger of homographs: in the browser, in domain name registries, and via dedicated testing tools.
- Dennis introduced the initiative of Universal Acceptance. Its aim is to foster the end-to-end vision with contributions from a wide range of stakeholders.
- Universal acceptance will be achieved once IDNs are available for everybody and in every tool: in web browsers, email clients, mobile applications, and user settings.

Related links: IRC • Video

In the Q&A session, the language technology efforts around Welsh were discussed. Keyboard layout for Welsh and treatment of diagraph characters have been an issue but this now seems to be resolved. Coding clubs are a means to gather interest from young people to work with the Welsh language. Components being developed for Welsh have the potential to be re-used for other small languages, e.g. in the area of speech and machine translation system development.

The multilingual cultural heritage data sets that have been developed for the JocondeLab portal may find usage also in commercial applications. The data is available under suitable licenses. One challenge during the data set creation process was granularity: e.g. DBpedia provides less detailed in some areas than certain culture heritage metadata.

Finally, the success of IDNs and the usefulness of networking around the topic in the last years were discussed. The effort of Universal Acceptance is a means to get important stakeholders involved. Events like the MultilingualWeb workshop series are essential to keep the conversation across stakeholder groups going. Dennis words fit well for closing the 8th MultilingualWeb workshop: It is all about meeting the right people and about keeping the conversation going.