D6.1.2: SUMMARY REPORT 1

Arle Lommel (DFKI), Felix Sasaki (DFKI)

Distribution: Public

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

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	6.1.2
Deliverable title:	Summary Report 1
Dissemination level:	PU
Contractual date of delivery:	31st July 2012
Actual date of delivery:	31st July 2012
Author(s):	Arle Lommel, Felix Sasaki
Participants:	DFKI
Internal Reviewer:	-
Workpackage:	WP6
Task Responsible:	Arle Lommel
Workpackage Leader:	Arle Lommel

The workshop report has been published online as HTML at

http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-workshop-report

This document contains a PDF version of the HTML page.

W3C Workshop Report: The Multilingual Web – Linked Open Data and MultilingualWeb-LT Requirements

11 - 13 June 2012, Dublin

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

The MultilingualWeb initiative was established to look at best practices and standards related to all aspects of creating, localizing and deploying the Web multilingually. It aims to raise the visibility of existing best practices and standards and identify gaps. The core vehicle for this is a series of events which started in 2010. Currently these workshops are supported by the MultilingualWeb-LT (MLW-LT) project. MLW-LT aims at defining ITS 2.0, which provides metadata for multilingual content creation, localization, and language technologies.

On 11–13 June 2012, the W3C ran the fifth workshop in the MultilingualWeb series in Dublin, “The Multilingual Web - Linked Open Data and MultilingualWeb-LT Requirements.” It was hosted by Trinity College Dublin (TCD) and Professor Vincent Wade, Director of the Intelligent Systems Laboratory at TCD, gave a brief welcome address.

The purpose of this workshop was two-fold: first, to discuss the intersection between Linked Open Data and Multilingual Technologies (11 June), and second, to discuss requirements of the W3C MultilingualWeb-LT Working Group (12–13 June). This workshop was different than the previous workshops because it focused on two specific topics and targeted a smaller audience (to allow for more discussion). Future MultilingualWeb Workshops will return to the broader scope of previous workshops and aim to be a meeting place for content creators, localizers, tools developers, language technology experts, and others to discuss the challenges of the multilingual Web.

The Dublin event ran for three days. The final attendance count was 78, almost exactly the number the organizers had felt was ideal for having deeper discussions about the chosen topics. The audience and speakers encompassed key representatives from the EU, Europeana, the linked open data community, localizers working on linked open data, ISO TC 37, localization companies and research institutions working on terminology, the W3C, and also flagship EU projects in the realm of language technologies like META-NET. This was an ideal group of stakeholders working on the intersection between linked open data and the Multilingual Web.

As with previous workshops, the presenters were video-recorded with the assistance of VideoLectures, who have made the video available on the Web. We again made

live IRC scribing available to help people follow the workshop remotely, and assist participants in the workshop itself. As before, people Tweeted about the conference and the speakers during the event, and you can see these linked from the program page. (Note that video is not available for working sessions, which were highly interactive. Those interested in the content of these sessions should consult the IRC transcript for more details.)

After a short summary of key highlights and recommendations, this document provides a short summary of each talk accompanied by a selection of key messages in bulleted list form. Links are also provided to the IRC transcript (taken by scribes during the meeting), video recordings of the talk (where available), and the talk slides. Almost all talks lasted 15 minutes. Finally, there are summaries of the breakout session findings, most of which are provided by the participants themselves. It is strongly recommended to watch the videos, where available, as they provide much more detail than do these summaries.

Contents: Summary • Day 1: Linked Open Data • Days 2 & 3: MultilingualWeb-LT Requirements

Summary

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

Linked Open Data

The workshop opened with short welcome addresses from Vincent Wade, Richard Ishida, Dave Lewis and Kimmo Rossi. Kimmo Rossi emphasized that in the planning of the upcoming Horizon 2020 and other European research programs under consideration, language technology will be seen in the context of the data challenge, that is: linked open data, big data, the data cloud, etc. This workshop is a good opportunity to gather ideas about the relation between language technology

and data, including topics for upcoming EC funding opportunities.

The setting the stage session started with our keynote speaker, David Orban. He looked into the future of computing devices and the "Web of Things". This is both a technical revolution and a societal challenge. We need to address these both on a personal and a political, decision-making level. The keynote was followed by two presentations: Peter Schmitz presented developments within the European Publications Office related to multilingual linked open data, and Juliane Stiller and Marlies Olensky described the European digital library Europeana, which features core multilinguality in its user interface and data, and which is currently being converted to used linked open data. Both presentations were similar with respect to their perspective: currently, huge, multilingual data sets are being made available as linked open data. The presentations were different with regards to application scenarios of the data, including different user groups and consuming tools. For both data sets, the interlinkage with the current linked open data cloud is an important task.

The linking resources session concentrated on this aspect of interlinking information. The speakers provided three different perspectives on the topic:

- Sebastian Hellmann introduced requirements and outcomes needed for making linguistic resources available on the Web and linking them with other parts of linked open data. Dedicated standardization efforts like the creation of the NLP Interchange Format (NIF) are needed to make this interlinkage happen. The benefit can be the easier re-use of data and integration of tools for natural language processing on the Web.
- Dominic Jones provided a view on linked data from the perspective of localization. Here, modeling of processes like translation by humans and / or machines, pre- or postediting, quality control etc. currently suffers from a diversity of tools and data formats. Linked data and the related standardized APIs can help to solve this problem; but, different to areas like e.g. linguistic research, the openness of the data can be a barrier for adoption, since localization is an industry in which high quality data rarely will be made available for free.
- As the last speaker in the "linking resources" session, Jose Emilio Labra Gayo provided several best practices about how to achieve multilinguality. Unfortunately, an analysis of the current state of the data shows that creators of linked open data rarely follow such recommendations, e.g., natural language descriptions (so-called "labels") rarely exist for languages other than English. A first step towards solving this problem might be to document the guidelines. As

a concrete outcome of this discussion, the W3C Internationalization Activity has put the creation of such a best practice document on its agenda.

The main topic of the linked open data and the lexicon session was how to make use of existing language resources—primarily in the realm of lexicons and terminology within linked open data, and vice versa. This section featured five presentations:

- Alan Melby introduced the efforts around RDF-TBX, an approach to represent terminological data based on the widely used TermBase eXchange (TBX) format as linked open data, that is: RDF. The efforts around RDF-TBX have the potential to be a bridge building mechanism, both between technologies and communities.
- Ioannis Iakovidis demonstrated the general tendency of terminology systems to "go on the Web": both the terminology data itself and access to it in various workflows (machine translation, terminology management, quality control) is moving away from proprietary technologies and formats.
- Tatiana Gornostay continued this topic with a focus on what can be done with cloud-based platforms for handling language resources, including terminology, and integrating them with other technologies. A yet to be solved problem is the mismatch between the modeling capabilities that linked open data provides, and the modeling needed for lexical resources.
- The presentation from John McCrae introduced Lemon (Lexicon Model for Ontologies), which can help to resolve the mismatch identified by Gornostay.
- Phil Archer closed the "linked open data and the lexicon" session by bringing in an aspect of creating and linking information into the discussion that had not been mentioned so far: like with any other kind of data, linked open data, esp. from the area of eGovernment, is subject to national and cultural needs and differences. Even the creation of URIs (Universal Resource Identifiers), that is the basic building blocks of linked open data, needs to take these political challenges into account. Otherwise even technically well thought solutions will not be adopted widely.

In the last session of the first day, Georg Rehm introduced the Strategic Research Agenda (SRA) being developed within the META-NET project. The SRA is one main piece of input for research in the Horizon 2020 and other programs. The presentation of the SRA at the workshop and discussions with participants had a crucial influence on framing the SRA with regards to the relation between linked open data and

language technologies, see the SRA section “Linked Open Data and the Data Challenge”.

In addition to the above presentations, during the first day several longer discussion sessions were held: a working session about identifying users and use cases - matching data to users, and a final discussion of action plans.

MultilingualWeb-LT Requirements

The second and the third day of the workshop were dedicated to gathering requirements about the Internationalization Tag Set 2.0, which is currently being developed within the MultilingualWeb-LT W3C Working Group. Several presentations highlighted different requirements. These are summarized briefly below.

After a welcome session for the ITS 2.0 discussion, the discussion of representation formats for ITS 2.0 metadata started with a presentation by Maxime Lefrançois. He explained the challenges for making the metadata available for or via Semantic Web technologies like RDF.

Three working sessions followed the representation format discussion: a session on quality metadata, on terminology metadata, and on updating ITS 1.0. The outcomes of these sessions and the two workshop days about ITS 2.0 requirements are summarized at the end of this executive summary.

In the content authoring requirements session, two presentations were given. Alex Lik emphasized the need of assuring interoperability of ITS 2.0 metadata with content related standards like DITA or localization workflow related standards like XLIFF. Des Oates reiterated that requirement with a focus on large scale localization workflows within Adobe. In these not only metadata is needed, but also the possibility for metadata round-tripping and a mechanism to expose metadata capabilities in a service oriented architecture.

The session on requirements on ITS 2.0 for localization scenarios started with a presentation from Bryan Schnabel. As the chair of the XLIFF technical committee, Bryan gave insights into current discussions within the XLIFF TC which are likely to influence the relation between XLIFF and ITS 2.0, and potentially the adoption of ITS 2.0 in general.

The second day closed with a presentation from Mark Davis about latest BCP 47 developments. BCP 47 is the standard for language tags, and recently extensions to BCP 47 have been created that overlap with functionality needed for ITS 2.0 metadata. The session was helpful in starting a coordination in standardization activities, that is now being continued between the relevant W3C working groups.

On the third day, the deep discussion sessions on ITS 2.0 requirements continued in

various discussion sessions on implementation commitments, project information metadata, translation process metadata, provenance metadata, translation metadata, a summary of implementation commitments and a session on coordination and liaisons between the MultilingualWeb-LT group and other initiatives.

In the translation process metadata session, David Filip talked about the need to have ITS 2.0 metadata available in complex localization workflows, reiterating statements e.g. from Des Oates and Alex Lik.

The requirements gathering for ITS 2.0 was especially successful in terms of raising awareness for the upcoming metadata in various communities. These range from large cooperates with localize huge amounts of content on a daily basis, using both human and machine translation, to smaller localization technology providers who have their specific technological solutions.

Another main achievement was the consolidation of requirements. The Requirements for Internationalization Tag Set (ITS) 2.0 had been published only a few weeks before the workshop and provided a huge amount of proposed metadata items. Via the discussions during the workshop, many proposals were consolidated, taking both needs in real life use cases and judgement of efforts into account. Follow the link to the first working draft of ITS 2.0 to see the set of metadata items agreed upon by workshop participants.

The work within the MultilingualWeb-LT working group will now focus on selected additional metadata and implementations of the metadata proposals with real life use cases, including various “real” client-company scenarios. In this way, the usefulness of the metadata for filling gaps in the way towards a truly multilingual Web can be assured.

Day One: Linked Open Data

Welcome session

Vincent Wade, Director of the Intelligent Systems Laboratory at TCD, welcomed the participants to Dublin and gave a brief speech emphasizing the important role that multilinguality and linked open data plays in Ireland, within CNGL or DERI, with support both from national and European funding.

Related links: IRC

Richard Ishida, W3C Internationalization Activity Lead and the leader of the EU MultilingualWeb project that funded the MultilingualWeb workshop series between 2009 and early 2012, emphasized the success of the MultilingualWeb workshop series so far and the plans to continue this series with a general MultilingualWeb workshop to be held next year.

Related links: IRC

Dave Lewis, research lecturer at the School of Computer Sciences & Statistics at Trinity College Dublin, introduced the participants to the goal of the first day: to discuss the intersection between multilingual technologies and linked open data, involving perspectives from areas like language technology, localization, and the Web. The 7th framework program of the EU, the upcoming Horizon 2020 program and national funding play an important role for moving these topics forward.

Related links: Slides • IRC

Kimmo Rossi the European Commission, DG for Communications Networks, Content and Technology (CONNECT), project Officer for the MultilingualWeb and the MultilingualWeb-LT projects, started with information about a reorganization within the EU: As of 1st July 2012, the European Commission Directorate General for Communications Networks, Content and Technology (DG

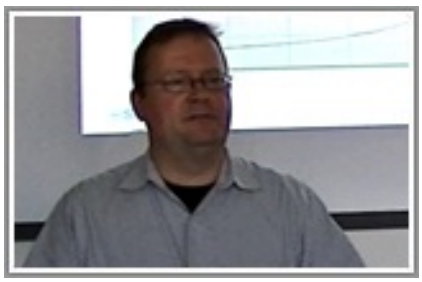
Related links: Slides • IRC • Video

CONNECT) has been created. Within DG CONNECT, the language technology area is now part of the data value chain unit: In the future, language technology will be seen in the context of topics like linked open data or public sector information. Language technology can help to leverage linked open data e.g. by extracting meaning from text or by converting structured data from unstructured data, and by bringing in work on language related resources (terminologies, ontologies, taxonomies etc.). Other significant remarks:

- Three EU calls are to be published in July, focusing on content analytics and language technologies, scalable data analytics, and an SME initiative on analytics.
- The EU currently is working out plans for a new funding program called Connecting Europe Facility (CEF).
- The CEF proposal currently encompass the digital library Europeana, eGovernment & eHealth, open data and multilingual access to online services.
- The Dublin workshop gathered speakers from many of these areas and hence was an important opportunity to provide input for shaping CEF.

Setting the Stage

The Setting the Stage session was chaired by **Paul Buitelaar** from the Digital Enterprise Research Institute (DERI).




David Orban, CEO of dotSub, gave the keynote speech, entitled “The

Related links: Slides • IRC • Video

Privilege and Responsibility of Personal and Social Freedom in a World of Autonomous Machines." In his talk he provided a look into the future at the next generations of computing devices. The Web and the physical world are coming together in the Web or "Internet of Things." What will be the impact of this development on our society, and how can we prepare individuals, and society as a whole to cope with the onslaught of accelerating change in our lives and our civilization?. David Orban provided some views about and answers to these questions in his keynote. The main points are summarized via the bullet points below.

- In the age of autonomous machines (computers, vacuum cleaners, cars…) is coming and we need to be prepared for it.
- When building the multilingual Semantic web, we need to be aware that machines will be the producers and consumers of much of our content, enabling them to perform functions that currently require human intervention.
- The human component in Wikipedia-like approaches, combined with large-scale semantic processing, will lead to a massive, hybrid, and powerful system.
- Social interactions including the Internet of Things are emerging, both on the personal and political, decision-making level.
- David Orban proposed to follow the "Proactionary Principle": With the new technological developments, we need to develop a balanced approach towards decision making.

Peter Schmitz, Head of Unit "Enterprise Architecture" at the Publications Office of the European Commission, talked about "Multilingualism and Linked Open Data in the EU Open Data Portal and Other Projects of the Publications Office." The talk presented contributions of the EU Publications Office (PO) to the Multilingual Web: daily publications in 23 languages, the multilingual thesaurus EuroVoc, multilingual controlled vocabularies, linked multilingual Web content, etc. The Publications Office is also heavily engaged in the European Open Data Portal and plans to provide dissemination of metadata as RDF, including persistant URIs. Other significant remarks:

Related links: Slides • IRC • Video

- Annotation of official content by users is problematic, because use cases, the notion of quality or modes of collaboration between online communities, and public services are hard to define.
- Provenance tracking is an unsolved problem for linked open data, yet it is

needed, e.g., for release management of EuroVoc or preservation of mappings in controlled vocabularies.
- Applications of linked open data in, e.g., eGovernment rely on sustainable, standardized identifiers like the “European Legislation Identifier.”
- For the MultilingualWeb-LT effort, multilingual linked open data resources like thesauri or authority tables can help to improve machine translation or other multilingual technologies.

Juliane Stiller and **Marlies Olensky**, researchers at the Berlin School of Library and Information Science at Humboldt-University Berlin, talked about “Europeana: A Multilingual Trailblazer”. Europeana aggregates content from 33 different countries to provide access to cultural heritage content across languages. Multilinguality in Europeana is implemented via the translation of the user interface, cross-lingual search, and subject browsing. It relies on translation of metadata to allow searches to cross language barriers and leverage content from one language to assist search in another. The Europeana Data Model (EDM) has been created to foster interoperability on a semantic level, and has already enabled to project to publish 3.5 million objects as linked open data. Other significant remarks:

Related links: Slides • IRC • Video

- EDM is available as a stable version, approved by several community workshops, encompassing specialists from libraries, archives, and museums.
- In the EuropeanaConnect project, alignment of Europeana with other resources like geonames has been achieved.
- Europeana encourages the re-use of its data by other European initiatives.
- Europeana is a trailblazer for the planned European Open Data Portal.

The Setting the Stage session on the first day ended with a **Q&A** period with questions and comments about the current and future state of multilingual technologies on the Web, about content from Europeana and the role of languages resources, and about the danger of “killing new topics by too much standardization.” For more details, see the related links.

Related links: IRC

Linking Resources

The Linking Resources session was chaired by **Dave Lewis** from Trinity College Dublin.


Sebastian Hellmann, researcher at the Agile Knowledge Engineering and Semantic Web research group at University of Leipzig, talked about "Linked Data in Linguistics for NLP and Web Annotation." The talk provided an overview of linguistic resources that are being added to the linked open data cloud, such as data extracted from language-specific Wikipedia editions and from the languages covered in Wiktionary, and the data developed within the Working Group for Open Linguistic Data. The Working Group is an Open Knowledge Foundation group with the goal of converting open linguistics data sets to RDF and interlinking them. For the sustainability of such resources, maintenance of the RDF data and open licences are important. Other significant remarks:

Related links: Slides • IRC • Video

- Ontologies provide a formal documentation for tools in natural language processing.
- The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between such tools, language resources and annotations.
- After conversion to NIF, a wide range of RDF tools can be used, for example, those developed within the LOD2 project.
- NIF version 2.0 will foster interoperability with the resources mentioned above as well as major RDF vocabularies such as OLiA, Lemon, and NERD.

Dominic Jones research assistant at Trinity College Dublin's School of

Related links: Slides • IRC • Video

Computer Sciences & Statistics, talked about “Linking Localisation and Language Resources.” The Drupal-based platform CMS-LION uses linked data as a way to connect content creators, consumers, language service providers, and translators. The output of various processes like text segmentation, machine translation, crowd-sourced post-editing, etc., is inter-connected through linked data, which is enriched using provenance data. Other significant remarks:

- Linked data allows for post-edits to be extracted and used in the re-training of Statistical Machine Translation (SMT) engines.
- Future plans for CMS-LION includes re-training of SMT systems with RDF based provenance information. This will be done in close cooperation with the PANACEA project.
- CMS-LION will also integrate with the SOLAS system developed at the University of Limerick.
- Finally, the Internationalization Tag Set (ITS) 2.0 metadata currently being developed in the W3C MultilingualWeb-LT working group will be processed within CMS-LION for XLIFF round-tripping scenarios.

Jose Emilio Labra Gayo,

Related links: Slides • IRC • Video

associate professor at the University of Oviedo, talked about “Best Practices for Multilingual Linked Open Data.” There are many projects that publish linked open data, such as data with labels/comments and other human-readable information, in various languages. Some issues occur frequently with multilingual linked open data; the presentation proposed seven best practices how to deal with these issues, related to e.g. the usage of Internationalized Resource Identifiers or multilingual vocabularies. Other significant remarks:

- URIs, that is resource identifiers, should be language agnostic; language specific information should be given by different labels and not by separate URIs.
- One should define human-readable labels for all URIs, but provide only one preferred label.
- Mechanisms that are also important for the Web like language tags or content negotiation should be used to make processing of information in different languages possible.
- Vocabularies in the Semantic Web should be multilingual; with this requirement, localization of ontologies becomes its own topic.

The Linking Resources session on the first day ended with a **Q&A** period with questions and comments about the usage of Internationalized Resource Identifiers within linked open data, difficulties with modeling lexical mismatches between languages (e.g. “Gift” in German vs. “gift” in English), various concepts for provenance and RDF based linguistic data in general. For more details, see the related links.

Related links: IRC

Linked Open Data and the Lexicon

The Linked Open Data and the Lexicon session was chaired by **Arle Lommel** from the German Research Center for Artificial Intelligence (DFKI).





Alan Melby, director of the Translation Research Group at Brigham Young University, gave a talk about “Bringing Terminology to Linked Data through TBX.” Terminology and linked open data are currently separate fields. To address this challenge, the TermBase eXchange (TBX) format in its latest version is currently being developed to include an isomorphic RDF version, RDF-TBX. This will allow to integrate the huge amount of linked open data available into terminological applications. For the linked open data community, TBX based resources can help

Related links: Slides • IRC • Video

with disambiguation tasks in multilingual scenarios and a concept based approach towards translation of linked open data. Other significant remarks:

- Via an RDF representation of TBX, linked open data can be enriched with links to a termbase that has been converted to RDF-TBX.
- Using conceptual, that is language agnostic resources available in the Semantic Web, an automated conversions between existing terminology resources may be possible.
- RDF-TBX may be able to address the separation between terminology and linked open data; this workshop was helpful in building the necessary human relations between the two communities to make this happen.

Ioannis Iakovidis, managing director at Interverbum Technology, gave a talk about "Challenges with Linked Data and Terminology." The integration of structured terminology data into linked open data is a challenge, especially in commercial environments. The TermWeb solution is a Web-based terminology management system that integrates with widely adopted software, e.g., office applications. There are various challenges that needs to be addressed for Web based terminology management systems, e.g.:

Related links: Slides • IRC • Video

- Tools use different APIs based on different technology stacks, which makes integration hard.
- Term identification is not only a task on the Web but also within various types of content such as software code or different content formats.
- The context of a term needs to be taken into account, again both on the Web and in other areas.

Standardization in various areas is a prerequisite to address these challenges. The formats and technologies provided by linked open data may be the standardized technology stack that provides the functionality needed to achieve these goals.

Tatiana Gornostay, terminology service manager at Tilde, talked about "Extending the Use of Web-Based Terminology Services." Tilde is currently advancing the establishment of cloud-based platforms for acquiring, sharing, and reusing language resources. The main application scenario is to improve automated natural language data processing tasks like machine translation. The exposure of terminology services on the Web allows for integration with machine translation

Related links: Slides • IRC • Video

systems, indexing systems, and search engines. Other significant remarks:

- Terminology is a language, “spoken” by language workers: translators, terminologists, technical writers, etc.
- Terminology can bridge the communities of linked open data, multilingual Web, and multilingual language technologies.
- EuroTermBank is a current solution for sharing and exchanging terminology, via the META-SHARE repositories of language data.
- Projects related to terminology on the Web include TaaS (Terminology as a Service), Let's MT!, ACCURAT and TTC.

John McCrae, research associate at the University of Bielefeld, talked about “The Need for Lexicalization of Linked Data.” The talk presented Lemon (Lexicon Model for Ontologies), a model for describing lexical information relative to ontologies. Lemon is based on existing models for representing lexical information (Lexical Markup Framework and SKOS) and aims to bridge the gap between the existing linked data cloud and the growing linguistic linked data cloud. Via the linguistic linked data cloud, a significant amount of multilingual data can be re-used in a variety of applications. Other significant remarks:

Related links: Slides • IRC • Video

- While linked data is frequently independent of language, the interpretation of this data requires natural language identifiers in order to be meaningful to the end user.
- Labels in linked open data are not used frequently; multilingual labels make up less than one percent of the linked open data cloud.
- Applications of LEMON include for example answering natural language questions about linked data, adaptation of linked data vocabularies to machine translation in new languages, and natural language generation.

Phil Archer, W3C team member working on eGovernment, talked about “Cool URIs Are Human Readable.” The EU is developing various standardized vocabularies for describing people, businesses, locations, etc. Currently the documentation for these is in English; in addition to the task of finding funding for the localization effort, political issues come into play: language, cultural identity, and trust need to be taken into account. This requirements influences basic decisions like choosing an URI that is neutral and acceptable across national borders:

Related links: Slides • IRC • Video

Vocabularies must be available at a stable URI, be subject to an identifiable policy on change control, and be published on a domain that is both politically and geographically neutral. Other significant remarks:

- One must not overlook the importance of branding and/or national identify inherent in a domain name.
- A domain name like `xmlns.com` is geographically and politically neutral when compared to `xmlns.eu`.
- Localisation of linked open data via different URIs for different language are not a good practice, but may actually help to increase adoption and interoperability.

Related links: IRC

The Linking Open Data and the Lexicon session on the first day ended with a **Q&A** period with a discussion about how the problems discussed relate to linguistic research and modeling in the area of semantics (which has a long tradition), about the limits of formalizing concepts, and about issues with the translating terms in different languages. For more details, see the related links.

Identifying Users and Use Cases Matching Data to Users

The Identifying Users and Use Cases Matching Data to Users session was chaired by **Thierry Declerck** from German Research Center for Artificial Intelligence (DFKI).

Related links: IRC

This working session featured discussion about the use cases and user groups interested in linked open data. The process of identifying users and getting their feedback on requirements is difficult, particularly because, while some users are multilingual, most are monolingual and gaining their input requires careful analysis of server logs

and other resources to determine how they interact with content. One of the difficulties for multilingual linked open data is that it is often generated at the end of the production chain, when it is difficult to make needed changes. Finally, it was emphasized that developers need to recall that linked data does not always equal linked *open* data: there are many cases where users will want to use linked data technologies but cannot make data open due to legal or privacy restrictions (e.g., linked data refers to identifiable personal information). As a result linked open data developers need to consider security requirements and build in ways to deal with them early on, and consider the ways in which linked data can move between open and closed categories.

Building Bridges

The Building Bridges session was chaired by **Kimmo Rossi** from the European Commission.

Related links: Slides • IRC • Video

Georg Rehm from DFKI gave a talk about the "META-NET Strategic Research Agenda and Linked Open Data." META-NET is developing a Strategic Research Agenda (SRA) for the European Language Technology research community. For the SRA, three priority themes have emerged: (1) the Translation Cloud, (2) Social Intelligence and e-Participation, and (3) Socially Aware Interactive Assistants. In all these areas, linked open data plays a crucial role. For example, data can be used to generate cross-lingual references between named entities as a part of the translation cloud, or to exploit publicly available government data to foster e-participation across Europe. Other significant remarks:

- As other presentations in the workshop have shown, interlinked language resources are becoming an inherent part of the linked open data cloud.
- This development shows that the linked open data community and the language technology community should work together in developing further long-term research goals.

- The “data challenge” will be an important aspect of the upcoming Horizon 2020 research program.

Related links: IRC

The Building Bridges session on the first day ended with a **Q&A** period with questions and comments about the legal frameworks discussed in the session, the prospects for disruptive change from language technology and how far things have come in just the past decade, and the difficulties in extending human language technologies for smaller languages, and other issues around extending multilingualism’s benefits to those in need. For more details, see the related links.

Action Plans

The Action Plans session was chaired by **Dave Lewis** from Trinity College Dublin.

Related links: IRC

Dave Lewis led the discussion about action plans coming out of the first workshops day. The discussion focused on what is needed to achieve concrete developments in the area of multilingual linked open data. As a community, practitioners need to address motivations and incentives for maintenance: as present much of the data is made available by researchers to support specific projects, but extending data and maintaining it is an expensive proposition and the community needs to figure out sustainable models to address these issues. There is also a clear need for best practices to help organizations make their data available and lower the costs involved in doing so. One issue is that those who benefit from linked open data may or may not be the same people who bear the cost of publishing it, so we do not yet see the business model needed to support extensive issues in this area, with a focus on resolving payment issues. Kimmo Rossi closed my stating that he sees these developments as taking a number of years, but that we need to start reaching out now to the broader linked open data community to ensure that multilingualism is a

core concern and that the solutions that are developed will meet the needs of a multilingual society.

Days 2 & 3: MultilingualWeb-LT Requirements

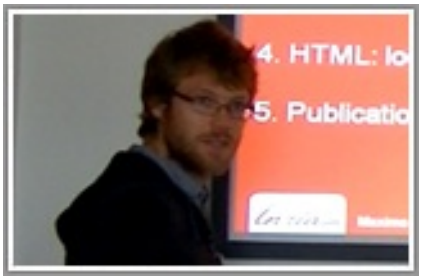
Welcome Session

Felix Sasaki, senior researcher at DFKI and W3C fellow, talked about “State of Requirements.” He introduced the Requirements for Internationalization Tag Set (ITS) 2.0 Working Draft and discussed its contents and the process for moving data categories forward in the W3C process.

Related links: IRC

Representation Formats: HTML, XML, RDFa, etc...

The Representation Formats session was chaired by **Jirka Kosek** from University of Economics Prague.



Maxime Lefrançois, researcher at Inria, talked about “MLW-LT, the Semantic Web, and Linked Open Data.” The standardization efforts of the MultilingualWeb-LT working group is focused on HTML5 and XML as content formats. The challenge is to integrate these formats and ITS 2.0 metadata with

Related links: Slides • IRC • Video

Semantic Web and linked open data approaches. Addressing the challenge will help to interface better with the linked open data community. Other significant remarks:

- Various technological approaches to add metadata to HTML5 or XML exist: RDFa, microdata, NIF (esp. for linguistic annotations), and general attributes in HTML5 and XML, used also in ITS 2.0.
- A main question is how to map XML specific mechanisms in ITS (e.g. so-called global rules) to the RDF world.
- Other ongoing standardization work like the development of the W3C provenance model can provide to a potential solution.

Quality Metadata

The Quality metadata session was chaired by **Phil Ritchie** from VistaTEC.

Related links: Slides • IRC

Phil Ritchie of VistTEC was joined by Arle Lommel (DFKI) to discuss a proposal for an ITS 2.0 model for marking language quality data in XML and HTML5 documents. The proposal sparked discussion about the needs of the group, the difficulties in exchanging quality data, and the verbosity of the proposed model (with standoff markup as one proposal for how to address the issue of verbosity and also data security). Overall the group felt that the proposal needed more elaboration in some key points and Ritchie and Lommel were charged with moving it forward.

Terminology Metadata

The “Terminology metadata” session was chaired by **Tadej Štajner** from Jožef Stefan Institute.

Tadej Štajner led a discussion about the ways to address terminology needs in ITS 2.0. While fully automated terminology work is not possible, there is a need for mechanism to tag term candidates in documents so that subsequent processes can evaluate them and make appropriate decisions. In discussion, the following points were raised:

Related links: Slides • IRC

- While ITS 1.0 already addressed term definition, there is a need for more sophisticated metadata about terms to support the full lifecycle of terminology research.
- The relationship between human- and machine-oriented terminology needs to be addressed. The requirements for each are different, and any solution must account for these differences.
- The development of an ITS 2.0 markup for terminology holds great promise to help content creators deal more effectively with terminology and integrate more closely with their partners.

Updating ITS 1.0 (Review of Data Categories from ITS 1.0)

The “Updating ITS 1.0” session was chaired by **Felix Sasaki** from DFKI / W3C.

In this section the group discussed updates to the existing ITS 1.0 data categories in ITS 2.0 to ensure that

Related links: IRC

there was agreement about the details. In most cases,
existing data categories are being adopted as is, but there was particular discussion about updating the ruby model to take into account work done in the HTML5 ruby model.

Content Authoring Requirements

The Content Authoring Requirements session was chaired by **Moritz Hellwig** from Cocomore.



Alex Lik, localization manager and information designer at Biosense Webster, talked about "CMS-Based Localisation Management." In topic based, technical authoring with standards like DITA, quite often many regulations need to be taken into account, e.g., the various ISO standards specific to the underlying domain. This adds a level of complexity to challenges, such as the variety of authoring platforms, scattered authoring teams, and the general information architecture. There is a strong need for organizing such complex workflows, and metadata is crucial for this. Nevertheless the metadata needs to fulfill various expectations to achieve wide spread adoption. Other significant remarks:

Related links: Slides • IRC • Video

- Interoperability with standards like DITA or XLIFF is a key requirements for ITS 2.0 metadata.
- Implementation of the metadata needs to be easy, and using the metadata should "show up on the bill."
- Education about the benefits and caveats of metadata and basic metadata

usage is needed for authors and especially for information architects.

Des Oates, localization solutions architect at Adobe, talked about "Adobe's Content Localisation Process." For a company like Adobe with large volumes of global content, the localization process is a complex workflow with many systems and services being employed. ITS 2.0 metadata can help to reduce the impedance across services, if it fulfills certain requirements. One requirement is that the metadata *survives* in a service oriented architecture with many systems, disparate technologies, and multiple technology vendors. Metadata round-tripping with XLIFF is key to making this happen. Other signifiant remarks:

Related links: Slides • IRC • Video

- Content goes through many inputs and outputs: authoring, terminology checking, machine translation service, publication etc. In Adobe, a mediation layer keeps these as services together; the services are potential consumers or providers of ITS 2.0 metadata.
- When information travels through workflows without a standard form of metadata, loss of information is inevitable.
- ITS 2.0 metadata is a connection mechanism between content creation, localization and publication. This is an important message to demonstrate the value of ITS 2.0 to decision makers in the related industries and departments.
- It is crucial that ITS 2.0 not try to define services themselves, but rather focus on defining the interfaces between services, keeping an agnostic stance about how jobs are done, but rather dealing with how they relate and pass data.

Localization Requirements

The Localization Requirements session was chaired by **Yves Savourel** from Enlaso.

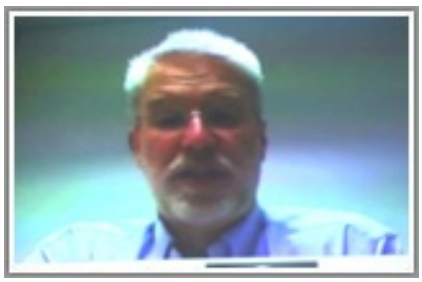
Bryan Schnabel, chair of the XLIFF Technical Committee, gave a speech, entitled “Encoding ITS 2.0 Metadata to Facilitate an XLIFF Roundtrip.” Currently XLIFF 2.0 is under development and one major topic under discussion is how to support extensibility in XLIFF 2.0. The decisions about extensibility will also influence the role ITS 2.0 metadata may play in XLIFF. The presentation discusses three options with regards to extensibility in detail and shows how to achieve metadata round-tripping with XLIFF and various content formats. Other significant remarks:

Related links: Slides • IRC • Video

- XSLT-based roundtripping of content formats is available for DITA and the Drupal CMS.
- One way to move forward with ITS 2.0 metadata in XLIFF 2.0 would be to provide native support in XLIFF 2.0 rather than not to address ITS 2.0 as an extension.
- Communities outside the XLIFF technical committee should also raise their voice in the extensibility discussion, to assure interoperability with various content formats and metadata items.

BCP 47 Developments

The BCP 47 Developments session was chaired by **Felix Sasaki** from DFKI / W3C.



Mark Davis, Sr. internationalization architect at Google and president of the Unicode Consortium, talked about “Coordinating the BCP47 ‘t’ Extension with MLW-LT Data Categories.” The standard for language tags BCP 47 currently has two so-called registered extensions (see the language tag extensions registry): one for setting behavior in locale APIs, the Unicode Locale Extension (“u”), and an extension (“t”) to identify content that has been transformed, including but not limited to:

Related links: Slides • IRC • Video

transliteration, transcription, and translation. Since ITS 2.0 metadata is also concerned with these processes, coordination is needed, to avoid conflicts between metadata conveyed via the BCP 47 extension and the to-be-developed ITS 2.0 metadata. Other significant remarks:

- The BCP 47 “t” extension for transformations can be used both for identifying that a transformation has happened (e.g., text has been translated from English to Russian), and for requesting a translation e.g. from a Web service (e.g., “translate this text from English to Russian”).
- Various types of transformation are registered in the Unicode Common Locale Data Repository, such as general transformations and machine translation.
- Tools harvesting Web content, e.g., for machine translation training, need to do the right thing for processing language tags with or without the “t” extension.
- The “t” extension is not meant for usage in localization-specific structured data like XLIFF.

Implementation Commitments

The “implementation commitments” session was chaired by **Felix Sasaki** from DFKI / W3C.

Related links: IRC

The session discussed the importance in the W3C process of getting firm implementation commitments very soon since any data category without implementation commitments will be dropped. The current status of these commitments will be maintained on a commitment wiki.

Project Information Metadata

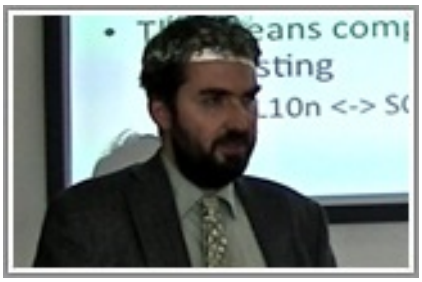
The Project Information Metadata session was chaired by **David Filip** from University of Limerick.

Related links: IRC

This session discussed various proposal for project information metadata. It resulted in dropping translation *qualification*, *genre*, *format type*, *purpose*, and a number of other metadata proposals, but retained *domain* for discussion, although a number of issues need to be resolved.

Translation Process Metadata

The Translation Process Metadata session was chaired by **Pedro Díez** from Linguaserve.



Related links: Slides • IRC • Video

David Filip, CNGL, talked about "Translation (Localization) Process Metadata?." All ITS 2.0 metadata items are orthogonal. However, in enterprise environments, the metadata has to work with many different types of workflows (see also the presentation from Des Oates) in a huge, service-oriented architecture. This includes, for example, terminology and translation memory life-cycles. In such a scenario, orthogonal categories must sync on the fly. Under the umbrella of the MultilingualWeb-LT working group, Trinity College Dublin and University of Limerick are working on implementations that help to support this requirement for ITS 2.0 metadata. Other significant remarks:

- Provenance models should not be invented separately for each metadata item. Instead, they should be specified using the same model in a consistent manner.
- The SOLAS system developed at the University of Limerick is the test bed for a service oriented architecture using metadata in localization workflows.
- The current set of ITS 2.0 metadata includes many proposals related to the translation and localization process; these need to be consolidated.

Related links: Slides • IRC

Pedro Díez led a discussion on various process-related metadata proposals with the goal of creating data-driven processes. Various categories like readiness indicator and state attributes could offer a lot of value to users, but the relationship with APIs needs to be clarified. ITS 2.0 also needs to be coordinated with XLIFF in this area to avoid a collision of definitions between the two. ITS 2.0 has the potential to bring together stages outside of the scope of XLIFF, such as CMS-side authoring.

Provenance Metadata

The Provenance Metadata session was chaired by **Dave Lewis** from Trinity College Dublin.

Related links: IRC

Provenance has emerged in recent years as a major topic because knowing where content came from and how it was produced can have a major impact on what is done with it, what quality processes are used, and how trustworthy the content is. The W3C currently has a Provenance working group, and it was agreed that the ITS 2.0 team should approach the Provenance working group to ensure that our efforts are coordinated with theirs. There remains considerable work to be done on defining the ITS 2.0 provenance model and specifying the use cases, but this topic is one that will continue to be very important.

Translation Metadata

The Translation Metadata session was chaired by **Yves Savourel** from Enlaso.

Related links: IRC

Much of the discussion in this section centered on the *target pointer* proposal, which has implementation commitments from Yves Savourel and Shaun McCance have both committed to implementing it because it allows for basic processing of multilingual XML files without requiring full support for an XML file in some cases. The second category was locale filter, which provides a way for authors to specify into which locales specific content items should be translated. The *auto language processing rule* category required more elaboration. Next the *term location* category was discussed as an equivalent to the XLIFF *restype* category. The end result was that more information is needed on some of these categories before decisions can be made whether to incorporate them or not.

Implementation Commitments and Plans

The Implementation Commitments and Plans session was chaired by **Felix Sasaki** from DFKI / W3C.

Related links: IRC

The goal of this section was to gain firm implementation

commitments for data categories. It was agreed that all implementation commitments should be finalized by mid-July, at which point Felix Sasaki was to create a new draft of the specification that included those categories for which there were sufficient commitments. All creators of data categories were to work to build consensus on their proposals and commitments to implement the proposed definitions.

Coordination and Liaison with Other Initiatives

The Coordination and Liaison with Other Initiatives session was chaired by **David Filip** from University of Limerick.

This working session discussed the relationship between the ITS 2.0 effort and other groups, such as XLIFF, ISO TC 37, and the ETSI ISG LIS, as well as participation in the upcoming FEISGILTT conference in October to be held in conjunction with Localization World. David Filip invited participants in the Workshop who were interested in participating in the program committee for that program to get involved.

Related links: IRC

Closing Remarks

The Fifth Multilingual Web Workshop was closed by **Arle Lommel** from DFKI.

This section ended the Workshop with special thanks to the staff of Trinity College Dublin, especially Eithne McCann and Dominic Jones for their support in running a successful workshop. The next Workshop, which will return to the broader, more general format of previous Multilingual Web workshops, was announced tentatively

Related links: IRC

for Rome in March 2013.

Authors: Arle Lommel, Felix Sasaki. Contributors: Nieves Sande and Richard Ishida as well as the scribes for the sessions: Pedro Díez, David Filip, Declan Groves, Moritz Hellwig, Milan Karasek, Jirka Kosek, Phil Ritchie, Yves Savourel, and Tadej Štajner. Photos courtesy Richard Ishida, Arle Lommel, and Leroy Finn. Collage photos by Arle Lommel and “themaxi” (sxc.hu). Videography by Leroy Finn. Hosting of video content by VideoLectures.