SEVENTH FRAMEWORK
PROGRAMME

LIDER: FP7 – 610782

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

Deliverable number	**D4.7**
Deliverable title	**Third Roadmapping Workshop Report**
Main Authors	**Bettina Klimek, Felix Sasaki**

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

Coordinator	Asunción Gómez-Pérez
Address	Campus de Montegancedo sn. 28660 Boadilla del Monte, Madrid, Spain
Reply to	asun@fi.upm.es
Phone	+34-91-336-7417
Fax	+34-91-3524819

Document Identifier	D4.7
Class Deliverable	LIDER EU-ICT-2013-610782
Version	1.1
Document due date	31 October 2014
Submitted	27 October 2014
Responsible	W3C/ERCIM
Reply to	fsasaki@w3.org
Document status	final
Nature	O(Other)
Dissemination level	PU(Public)
WP/Task responsible(s)	Felix Sasaki, DFKI / W3C Fellow
Contributors	-
Distribution List	Consortium Partners
Reviewers	Reviewed by the project consortium
Document Location	http://lider-project.eu/?q=doc/deliverables

Executive Summary

This document, the third roadmapping workshop report, summarizes the outcome of three roadmapping activities:

- The fourth LIDER roadmapping workshop, held 2 September with about 40 registered participants and in alignment with the Semantics conference in Leipzig.
- A session about “XML, Semantic Web and Content Analytics”, held with about 40-50 participants at the 2014 XML Prague conference.
- A presentation and discussion about linked data and technical documentation, held with about 40 participants at the 2014 SOAP! conference.

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

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

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

Document Information

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

Deliverable	**Number**	D4.7	**Title**	Third Roadmapping Workshop Report
Workpackage	**Number**	4	**Title**	Community building and dissemination

Date of Delivery	**Contractual**	31 October 2014	**Actual**	27 October 2014
Status	version 1.3		final ■	
Nature	prototype □ report □ dissemination ■			
Dissemination level	public ■ consortium □			

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

Abstract (for dissemination)	This document, the third roadmapping workshop report, summarizes the outcome of three roadmapping activities: • The fourth LIDER roadmapping workshop, held 2 September with about 40 registered participants and in alignment with the Semantics conference in Leipzig. • A session about “XML, Semantic Web and Content Analytics”, held with about 40-50 participants at the 2014 XML Prague conference. • A presentation and discussion about linked data and technical documentation, held with about 40 participants at the 2014 SOAP! conference.
Keywords	LIDER, roadmapping workshop, report

Version	Modification(s)	Date	Author(s)

01	First Draft	23/10/14	Bettina Klimek, INFAI; Felix Sasaki, DFKI / W3C Fellow
02	Review version	24/10/14	Paul Buitelaar, NUI Galway; Roberto Navigli, UNIROMA
03	Final version	29/10/14	Felix Sasaki

Project Consortium Information

Participants		Contact
Universidad Politécnica de Madrid	UPM	Asunción Gómez-Pérez Email: asun@fi.upm.es
The Provost, Fellows, Foundation Scholars & The Other Members of Board of The College of the Holy & Undivided Trinity of Queen Elizabeth near Dublin (Trinity College Dubl, Ireland)		David Lewis Email: dave.lewis@cs.tcd.ie
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)	DFKI	Felix Sasaki Email: felix.sasaki@dfki.de
National University of Ireland, Galway (NUI Galway, Ireland)	NUI Galway OÉ Gaillimh	Paul Buitelaar Email: paul.buitelaar@deri.org
Institut für Angewandte Informatik EV (INFAI, Germany)	UNIVERSITÄT LEIPZIG	Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Universität Bielefeld (UNIBI, Germany)	Universität Bielefeld	Philipp Cimiano Email: cimiano@cit-ec.uni-bielefeld.de
Universita degli Studi di Roma La Sapienza (UNIVERSITA DEGLI STU, Italy)	STVDIVM VRBIS	Roberto Navigli Email: navigli@di.uniroma1.it
GEIE ERCIM (ERCIM, France)	W3C®	Felix Sasaki Email: fsasaki@w3.org

Table of Contents

1 Introduction

This document, the third roadmapping workshop report, summarizes the outcome of three roadmapping activities:

- The fourth LIDER roadmapping workshop, held 2 September with about 40 registered and in alignment with the Semantics conference in Leipzig.
- A session about "XML, Semantic Web and Content Analytics", held with about 40-50 participants at the 2014 XML Prague conference.
- A presentation and discussion about linked data and technical documentation, held with about 40 participants at the 2014 SOAP! conference.

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

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

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

2 The 4th LIDER Roadmapping Workshop

2.1 *Introduction*

This report gives a summary of the 4th LIDER roadmapping workshop, which took place on 2nd September 2014 as part of the SEMANTiCS conference pre-program and the MLODE 2014. For more information, take a look the workshop program.

The main objective was to identify areas and tasks in content analytics where Linked Data & semantic technologies can contribute. Due to the numerous companies presenting their use cases, special input was gathered concerning special needs for companies, businesses and enterprises. In what follows, a summary of each talk will be presented.

2.2 *Contributions*

Welcome and Introduction

Philipp Cimiano (University of Bielefeld) opened the workshop. After a short introduction of the EU LIDER project he outlined the goals of this workshop. The main focus is set on the use and needs of linguistic linked open data for the business and industry sector.

Thus, the two main objectives for the workshop day were:

- Identification of areas and tasks in content analytics where Linked Data & semantic technologies can contribute.
- Gathering input from experts and stakeholders in the area of content analytics as a basis to define a European R&D and Innovation Roadmap for the European Commission that will help the EC to prioritize future R&D activities.

Tatiana Gornostay*:* Language Meets Knowledge in Digital Content Management

Tatiana Gornostay (Tilde) reported on closing the gap between language and knowledge in digital content management. The main goal is to bring innovation to the market for human professionals and machine users by working on improving communication between engines. Terminology management shall be opened to broader applications in content management which include but are not only limited to machine translation.

Hereby, one main challenge is seen in the terminology management which primarily bases concepts on linguistic expressions next to the existing knowledge. The speaker emphasized that terminology should not only be regarded in the context of language but also in the context of content management and enrichment. As a basis for that the creation of rich content which is multilingually and semantically linked to data is needed. The benefits of this concept based approach to enrich existing content will result in a higher quality of terminology that will finally lead to a saving of time and resources.

Ilan Kernerman: Generating Multilingual Lexicographic Resources

Ilan Kernerman (K Dictionaries Ltd, Tel Aviv) introduced K DICTIONARIES Ltd. which have a long tradition in the lexicographic field of dictionary development. He shared his experiences facing a transition from traditional dictionaries to multilingual datasets, data management and software engineering, architectures and design due to the increasing technological development. Today the K Dictionaries Ltd. resources comprise multilingual databases for over 20 major and some minor languages including linguistic information on morphology and pronunciation, lexicographic editorial tools and applications.

The main focus lies in the quality of the language data, hence, the data is first collected and edited manually by first language speakers to build monolingual datasets which are then extended and connected to form bi- and multilingual datasets via automatic translations. The main goal is to get from traditional lexicography value for applications such as machine translation, e-learning, word processing, text mining and search engines.

The use of linguistic linked open data is desired regarding its interconnectedness in nature and the vast amount of available language data. However, the integration of this data suffers from the mediocre quality of the automatically created content. The challenge is to arrive at automatically generated high quality content that can cope with the central problems of resolving the complex cross-linguistic relations that have rarely a 1:1 equivalence (for instance in compound words) as well as extending the few existing quality sensitive domains, e.g. education and healthcare which are even now interested in high quality linguistic data.

Heiko Ehrig: Resources! Resources! Resources!

Heiko Ehrig (Neofonie) introduced the company shortly. They shifted from developing search engines web and mobile application development and consulting, including interaction design, testing and data analytics. Neofonie developed a German text mining API that performs classification, keyword detection, entity detection, date detection, NER, and quotes (API key http://bit.ly/txtwerk). From their experience with NLP and linked data they point to the examination of the following issues:

- extension of entity types.
- building more individual customer lexica and sentiment detection.
- broaden LD and NLP for more languages than English.
- development of a gold standard of German (N)ER.
- discussion of standardized text mining API.
- support of open data and open licenses.

Mark Zöpfgen: Software-Supported Bibliographic Recording and Linked Data

Mark Zöpfgen (German National Library) presented their library activities in content extraction and semantic web. They maintain the National Bibliography, which contains all national print and electronic publications since 1913. They produce an authority file (called GMD “Gemeinsame Normdatei”) with metadata. Activities in content extraction and semantic web comprise several projects. In these they build an ontology for generating the data and which enables a multilingual access to subjects in order to make the German National Library internationally available.

Manual effort is also invested in providing high quality translations of the subject headings of the bibliographical records into English and French. So far, an Open Linked Data Service for spreading the data is available and downloadable in RDF format under creative commons zero license. The main goals of the German National Library comprise the following topics:

- constant improvement of the poor formal state of the bibliographical highly reliable data.
- building an integrated portal with search engine and linked data.

- integration of German bibliographical data into The European Library and finding standards for the provision in the linked data format.
- increase precision of multi-language term mappings under the assumption that there is rarely 1-1 matching.
- the motivation of external parties to work with RDF data and improve search possibilities.

Massimo Romanelli: Social Media Monitoring: from Sentiment to Intention

Massimo Romanelli (Attensity Europe GmbH) introduced the company which provides analytics for customer engagement by retrieving conversational information from social media platforms such as Google and Twitter. He presented the LARA (Listen, Analyze, Relate, and Act) paradigm. A complex enterprise solution suite (http://www.attensity.com/products/) has been developed. Thereby Attensity Q exploits external resources for existing classifications via linked data. Attensity Analyze then combines the social with the internal data using NLP tools and text analytics. Finally, Attensity Respond displays easy topic metrics which suggest what the customer might want. The main goal is to detect the customer's intention from his sentiment represented in the text and to be able to react accordingly.

Even though Attensity makes already successfully use of NLP engines, knowledge engineering (Lingware) and annotated documents, more resources are needed to expand the vertical domains to identify the intention of a user. Such resources do not only encompass more data but also a model that represents pragmatic implications and could be therefore used to create the correct query to a certain customer need.

Marc Egger: Text Analytics for Brand Research -Non-reactive Concept Mapping to Elicit Consumer Perception

Marc Egger (Insius) talked about brand research in the context of product development in companies. On the basis of text analytics for consumer social media content, concept maps for market research are developed. The aim is to find out what the consumer thinks about product, brands and general topics via NLP tools that detect, collect and analyze textual consumer content from the web. As an example, the work with the brand concept map was presented. Out of this map the customers' associations are turned into a network representation that is then analyzed according to the values i) strength, ii) favorability, iii) uniqueness and iv) patterns of thought. This analytics software which is used to elicit consumer perceptions could be improved with regard to the textual data processing in various aspects. These include:

- refine POS tagging and dependency parsing for written oral language such as forum posts for more accurate concept candidate detection.
- also cover intra-article topic relevance.
- face aggregation challenges such as spelling mistakes (burger = burgr), synonymous concepts (tasty burger = delicious burger).
- increase accuracy in ratings of topic relevance by providing high quality resources for German NER and better German anaphora resolution tools.

Alessio Bosca: Linked Data for Content Analytics in CELI

Alessio Bosca (CELI) presented how CELI is exploiting linked data. Their focus is on speech applications, semantic search, text analytics, opinion mining and social media intelligence. The core technology used encompasses language processes such as language identification, morphological analyses and semantic analysis. CELI exploits the linked data in the LOD cloud a) as a user by making use of for NER, and b) as a provider for internal use and for crafting RDF artifacts. Two projects were addressed: a book project for the digital humanities and the Homer project for multilingual interfaces to assessing data from different public administration. From the work with linked open data the the LOD cloud community is advised to put more emphasis on truly linking of the datasets.

With regard to the public sectors it is suggested that more data should be published as linked open data and that international standards should be used. The issue of publishing companies' linked data under an open license was also addressed. The speaker made the point that besides the resistance to sharing, because of valid competitive concerns, company data is generally over-fitted to their solutions and clients. In other words, companies need to be able to manage 'micro-domains' which are regarded as less useful in general. As a compromise it was suggested by the audience that companies should not answer the question why they do not publish their linked data, but what they could publish.

Oscar Muñoz: Content Analytics for Media Agencies

Oscar Muñoz (HAVAS Media) shortly introduced the company. HAVAS Media is an agency that offers market studies which extend traditional market survey with social media analysis. The presentation mainly focused on how relevant touch points between brands and the consumers can be established. Therefore, consumer profiles are created by gathering information on the awareness, evaluation, purchase intention and post-purchase experience (e.g. integrated sentiment analysis with UPM) of the user. This intelligent consumer profiling further includes time series analysis, e.g. event detection and explanation, of relations between social buzz and advertising pressure. Another method used is the social graph analysis which detects influences, brand ambassadors, detractors, viralisers and content propagation.

Due to the large volume of data sources (e.g. over thousands of brands) as well as their heterogeneity several challenges arise. These are summarized in the following questions:

- How is it possible to associate different data sources from social media, search engine marketing, customer date, site analytics, offline advertising and digital display advertising?
- Can we arrive at Big Linked Data integrating multiple heterogeneous and unstructured data sources at scale?
- How can we tackle the problem of the variety and velocity of data sources to decrease the integration costs which are incurred by the lack of social media formats?
- In what ways must a consumer connection platform be improved to enable a cross platform information tracking that is able to infer online and offline user behavior?
- How is high a high accuracy ensured by a rising complexity resulting of multilingual processing?

Andreas Nickel: Applicated Insights: Computational Linguistics and Semantic Analysis as Part of Business Workflows

Andreas Nickel (Ferret Go) reported on the challenges of content analytics in the Ferret Go startup business that exists since 2012. They are mainly dealing with media and textual resources such as articles, reviews and other online text and provide a structured content analysis for their customers. Especially heterogeneous clients, e.g. newspaper reader comments and social media are challenging, which is why some work is still done manually. The structure in non-structured data is discovered by applying computational linguistics. Four use cases have been presented:

1. Insights in community management via fast moderation which is a real-time analysis of reader's comments for bild.de.
2. Insights in opinion management by tracking the user's opinion and analyze customer feedback in unstructured hotel reviews.
3. Feedback dispatching for commerce and industry, e.g. customer relationship integration, workflow priority.
4. Deep content mining e.g. topic detection for long periods of time.

The conclusions Ferret Go could draw from previous work are:

- More accurate ways of automated content analytics must be found since the manual effort is too high and the the quality of crowdsourced results is questionable as well.
- Companies often do not know what to do with analytics - aid must be provided to help the clients then to decide how they can react.
- Still we cannot get 100% accuracy, so should be aware of that fact while simultaneously effort should be invested to arrive at 100% in the future.
- Content analytics could be facilitated if clients take up the advice to store only selectively potentially relevant data rather than saving everything.

Patrick Bunk: Setting them up for Failure – How Customer Expectations Collide with Economic Realities of Text Analytics

Patrick Bunk (uberMetrics) talked about customer expectations with regard to text analytics. He outlined the functions of internal and external data within companies. The former is used for knowledge management and business intelligence, whereas the latter is mainly taken as analytical basis for search engines and market intelligence. Working in the fields of (social) media monitoring and sentiment analysis uberMetrics reports that their clients have a mean of 500k articles per month.

Speaking from previous experiences it can be concluded that expectation gaps and varying quality over time and domains become recurring issues. Both are addressed by focussing on the economic realities, which means that expectation fulfillment and quality are strongly connected to the different pricings of the various analytical approaches: free for automated 70-80% accuracy, 1 Euro per article for manual work, tailor made solution by training a customer model and with costs of employing one person for one year and also crowd based tagging with costs of 0.05 Euro per article. Finally, with respect to the economic and quality aspects of text mining tasks the following suggestions have been proposed:

- coping with failure gracefully.
- focus on generalized solutions.
- testing algorithms on humanities majors.
- be aware of manual labor substitute.
- tailor-made mining is at a local maximum pre scalable product.

- automation through knowledge should be socially beneficial.

Dirk Goldhahn: Introduction to the German Wortschatz Project

Dirk Goldhahn (University of Leipzig, NLP group) was the only speaker presenting a linguistic dataset from the academic field. He introduced the the Leipzig Corpora Collection. The dataset comprises corpus-based full form monolingual dictionaries for more than 220 languages which comes with a variety of meta-data, e.g. word frequencies, POS tagging and co-occurrences. Furthermore, the corpora are enriched with statistical annotations such as POS, topics, word and co-occurrence frequencies. At the moment the NLP group is working on a conversion of their data into a linked data format. At the same time integration work of external sourced still needs to be done.

Michael Wetzel: Towards the Single Digital Market – Processing Knowledge, Independent from Language

Michael Wetzel (Coreon) focussed on on the management of language resources that can be used in different applications. Thereby it was discovered that knowledge is mainly accessible by multilingual data, hence, forcing it to stay in knowledge silos. The approach undertaken to open up the knowledge access is to discard the string driven search/access, because it has to fail given that one and the same object has multiple expressions. Rather, it has to be searched for the thing instead of the string! This can be achieved by a fusion of concepts and multilingual terminology. For this purpose Coreon has developed a knowledge software that establishes a knowledge map starting from a multilingual terminology list. The primary challenge that has to be tackled is the need to bridge the various format standards of TBX, SKOW and OWL.

2.3 *Key points of the Workshop*

Philipp Cimiano closed the workshop presenting a summary of the most discussed topics during the workshop. Overall, most participants agreed that the issues of creating more standards as well as ensuring working links within the LOD cloud should gain more emphasis in the linked data communities. Further topics that are regarded as central work orders for the LIDER project were identified by a significant number of participants. These are summarized as key points below:

- Sharing of linked data involving a cooperative data curation.
- Providing more resources for micro-domains that generalize and can be shared.
- Avoid knowledge silos by emphasizing more the linking both in communities and enterprises.
- Focus on more high-quality open data.
- Work on deeper analysis and more semantics to enable semantic search for things, rather than strings.
- Clarification of what accuracy rates of linked data analytics are reasonable for clients with high statistical result expectations.

3 Linked data and XML tooling: XML Prague 2014

3.1 *Introduction*

This report summarizes a session held at the XML Prague 2014 conference. The session was dedicated to the topic of XML, Semantic Web and Content Analytics. Technologies like XML and RDF are rarely discussed in the same context. Speaking practically, XML tools that process RDF are not yet common - but they do exist, see below for more information.

With this background, it came to a surprise that the session was crowded. Between 40-50 people attended and gave feedback on various aspects of using XML and Semantic Web technologies in content analytics applications. Key discussion points are summarized below.

The session title uses the term Semantic Web. During the session also the term Linked Data was used to refer to the ability to represent machine readable, interlinked information on the Web.

XML Prague is a conference series with a great variety of attendees. Many, but not all are "geeks": they are interested in real code and tool demonstrations. They also share a strong interest in XML, but in recent years, more and more technologies are being discussed at XML Prague. The 2014 edition of the conference had presentations also in the main program on Semantic Web, browser technologies, layout on the Web and many other topics.

In the session, a small set of slides helped to introduce the topic. During the oXygen Users Meetup at XML Prague, a demo showed how to integrate automatic entity annotation functionality into the oXygen XML editor.

3.2 *Contribution*

The main part of the content analytics session was an interactive discussion. Key points are summarized below.

Target Audiences: Who needs to know about Content Analytics?

A reoccurring question during the session was: what type of user actually needs to know about content analytics and linked data / semantics? Depending on the user in question, requirements for tooling and usage scenarios differ.

Developers of XML editing tools may want to add basic functionality to their tools, like the forehand mentioned semi-automatic entity annotation. An important usage scenario for such annotations is to provide context for content authors and translators: entity annotations can help them to disambiguate the meaning of a content item easily, which can safe time in translation processes.

Some people may be called content architects. They are not dealing with a single piece of content or document, but with larger volumes. Two types of content architects can be differentiated: people who set up the actual processing chain technically, for processing potentially thousands of documents; and people who add value to large document sets. The former may want to add functionality to tools that are used by the latter, e.g., a way to decide what documents need specific review before publication, or a way to categorize documents (semi) automatically.

The classical and manual counterpart of such categorization tasks is done by human topic indexers. With the fast amount of data to be processed today, this job does not scale anymore. The semi automatic approach of categorization seems to be promising for this group of people. It allows them to work with general or domain and project specific controlled vocabularies, and applying these to fast amounts of content. However, topic indexers are still missing the tooling to work in this manner without becoming a software programmer.

In general, content architects, both from the technical side as well from the manual or semi automatic content processing side, so far don't have knowledge about linked data or about the content analytics technologies. They also rarely know about the potentials of content analytics application scenarios. But things are starting to change. In the XML Prague main conference, a presentation from Charles Greer showed how RDF data can be stored and queried within a major XML data base. The challenge is now to educate data base users. They need to know both SPARQL and XQuery to be able to work with this solution.

Usability of Content Analytics Tools

This aspect brings up the next topic: usability of content analytics tools. The forehand entity annotation example has the advantage that a user does not need to know anything about RDF, linked data or the automatic annotation process. The functionality can be used in the WYSIWYG environment of an XML editor. However, in practice, many content producer even don't use such editors, but rely on word processors.

There are two approaches one could take from here: first, to enhance the usability of content analytics solutions, and second, to add these solutions into the tools commonly used by the content author (working with one document) or by content architects (working with large volumes). The previous section made clear that adding entity annotation to a authoring tool hits just the tip of the ice berg; many other parts of the content production tool chain that are handled or at least set up by content architects need to be adapted. Examples are CMS systems, publishing pipelines, automatic type setting tools, or content integration portals.

Various session participants pointed out that more and more publishing houses have started to look into (semi)automizing metadata creation. The term metadata here is used to describe any kind of content related information. In this sense, the outcome of a content analytics process leads to content enriched with various types of metadata. That enrichment may happen on a word, paragraph, document or even document collection level. And all kinds of metadata may benefit from manual refinement.

Workflows and Interplay between Content related Technologies

For real deployment of content analytics solutions, it is important to integrate these into the appropriate part(s) of the content production workflow. The forehand mentioned XML data base allows to integrate semantic information into the XML data itself, and use SPARQL and XQuery at the same time for querying. A blog post provides further information about how this works technically. In this scenario, it is assumed that the actual content creation is finished. The data base is then processed by the content architect or by an end user.

Many participants in the session pointed out that a workflow including content analytics processes needs to allow for human intervention before producing final results. The previous section gave examples why this is needed for content categorization. Such intervention has the potential to improve the content analytics processes itself. However, the session participants were not sure wether the algorithms used in current content analytics tools are able to incorporate such feedback loops.

The final workflow aspect discussed was related to snapshots of semantic interpretation. Depending on what static or dynamic semantic resources are used, the outcome of a content analytics process may differ. An example is the Wikipedia categorization of a tablet computer. The first version of the related Wikipedia page has been created in 2006. The current version of the page categorizes a tablet computer as a kind of mobile device. But the tablet computer definition itself and this kind of categorization are not available in pre 2006 Wikipedia data. For such reasons, a user of content analytics tools not only wants to be aware of the relevant semantic resources, but also needs to know their temporal dimension.

Data Aspects in Content Analytics Applications

Data Formats and Storage

The previous discussion on XQuery and SPARQL has touched upon the aspect of storing semantic information in various tools and various parts of the content production workflow. The RDF/XML syntax provides a standardized way to store RDF as XML and as part of XML content items. However, the previously mentioned XML data base does not use RDF/XML, but rather a proprietary approach of storing stets of RDF triples.

Another piece of information that may need (further) standardization is how to store results of content analytics processes. The previously mentioned outcome of entity annotation processes can be stored as ITS 2.0 Text Analysis information, see a related example. However, ITS 2.0 defines Text Analysis in a rather broad sense and does not provide fine grained information that may be specific to a certain type of content analytics process (opinion mining, sentiment analysis, document categorization etc.). Further standardization may be needed.

An important lesson to learn from ITS 2.0 is that information about content analytics is only useful if analytics processing tools information is available. One reason is that tools output is difficult to compare e.g. in terms of quality or auto generated confidence scores. ITS 2.0 provides a Tools Annotation mechanism to identify the tools involved in producing analytics or other kind of information.

Language, Content Domains and Data Sharing

Many participants of the session are working with multiple languages and topic domains on a daily basis. From this experience, they pointed out that there is a need not only for general semantic resources, but also resources specific to the domains and languages in question. Hence, there should be efforts for building high quality & curated domain specific and multilingual semantic resources.

Especially in the realm of public open data such resources are more and more being created. A key challenge is to find business models that demonstrate the value of data, and that encourage people from both the public and the private sector to share their data. Content analytics solutions bear the potential to become a catalyst for open data applications, but this has still be to be proved by example.

Education about Content Analytics and Linked Data

Overall, the session clearly showed that there is a huge difference in terms of knowledge about linked data and content analytics. In this respect, several participant pointed out that the BBC has great examples that demonstrate the general value of semantic information. This can be used as a basis to educate people not aware of content analytics and linked data at all. But more education is needed, especially for demonstrating the role of linked data in more complex content analytics applications like sentiment analysis or opinion mining.

The session promoted both the topics of content analytics and linked data. Given that deep knowledge in both areas cannot be expected when talking to end users, one may have to put specific efforts to raise awareness about making their relationship clear. And especially providing clear and simple answers to the question "Why should one use linked data for content analytics?" may help to foster industry adoption for linked data based content analytics applications.

3.3 *Key points of the session*

The outcome of the XML Prague session can be summarized as follows.

Various groups of users can be identified for linked data aware content analytics tooling, from an individual content author to a content architect and indexing specialists working with masses of content. For all these users the usability of content analytics tools are of high importance. Tooling needs to be available in the right part of the content production workflow. Addressing certain standardization challenges can help the interoperability of metadata produced by content analytics applications. These applications only add value for the end user if they are tailored to selected domains and languages. In that way content analytics also has the potential to become a booster for business models demonstrating the value of public open data.

4 Linked data and technical documentation: Soap! Conference 2014

4.1 *Introduction*

SOAP! is a conference for technical documentation that took place in Krakow, Poland. At the conference a presentation and discussion on linked data and technical documentation was held.

Technical documentation is a huge market and has huge conferences like Tekom]. The community is interested in Web technology developments. That is, one will find regularly presentations about HTML5 at Tekom - but the audience is mostly in a listening mode. The SOAP! conference is of interest beause it is much smaller than Tekom - 150 participants compared to several thousands at the annual Tekom conference - but also more interested in new trends.

4.2 *Contribution*

The presentation was about ITS 2.0 and linking content in the realm of technical documentation to external multilingual data sources. Most of the examples used data sources like DBpedia and wikidata. The interest in this usage scenario was huge. Like other content related areas, technical documentation suffers from more and more content being available on the Web. The industry needs to find out how to make a difference to competitors. Content with additional assets, like ITS 2.0 metadata to ease localization, and further metadata for contextualization, personalization, or search engine optimization, could become the differentiator.

Noz Urbina, co-author of a book on “content strategy”, was arguing in a similar way. His presentation was less technical, and on purpose he avoided the terms “semantic web” or “linked data”“ and rather talked about “metadata”.

After his talk he was asked from the audience “What you presented sounds like Tim Berners Lee Vision of the semantic web - why don’t you use the term?“ His answer was that high level business people who decide about budget for content workflow tools etc. see semantic web as an academic area. They will not invest money. But if one focuses on the functionality and talks about metadata in general, the chances are higher to get interest.

4.3 *Key points of the session*

Technical documentation may become an important application area for multilingual linked data. Studies discussed at the event that show that users who have a problem with a technical device, software etc., search on the Web and find the answer often in manual content. But since currently most of the content is not enriched with metadata, they search for keywords.

In other areas like asking for a location, the time the weather etc., question answering functionality has already found relevance and robustness for Web search. Having such functionality for technical documentation may be attractive for many industry areas, and may become a differentiator to competitors.

Technical documentation already today offers a lot of content that contains highly structured information. This information is lost in the process from authoring to Web publishing. Formats like DITA or DocBook come with XML to HTML conversions, but so far existing metadata is not transported properly. This is an opportunity for existing technical documentation content to become a value asset.