LIDER: FP7 – 610782

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

Deliverable number	**D.3.2.1**
Deliverable title	Roadmap for the use of linguistic linked data for content analytics – Phase I
Main Authors	Philipp Cimiano, Matthias Hartung, Sebastian Walter (UNIBI) Asunción Gómez Pérez, Guadalupe Aguado de Cea, Elena Montiel Ponsoda, Victor Rodríguez Doncel (UPM) Dave Lewis (TCD) Paul Buitelaar (NUIG) Roberto Navigli, Tiziano Flati (UNIROMA1)

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	+34-91-3367417
Fax	+34-91-3524819

Document Identifier	D3.2.1
Class Deliverable	LIDER EU-ICT-2013-610782
Version	2.0
Document due date	30.9.2014
Submitted	31.10.2014
Responsible	Philipp Cimiano, Bielefeld University
Reply to	cimiano@cit-ec.uni-bielefeld.de
Document status	final
Nature	Report
Dissemination level	Public
WP/Task responsible(s)	Philipp Cimiano, Bielefeld University
Contributors	Philipp Cimiano, Matthias Hartung, Sebastian Walter (UNIBI) Asunción Gómez Pérez, Guadalupe Aguado de Cea Elena Montiel Ponsoda, Victor Rodríguez Doncel (UPM**)** Dave Lewis (TCD) Paul Buitelaar (NUIG) Roberto Navigli, Tiziano Flati (UNIROMA1)
Distribution List	Consortium Partners
Reviewers	Felix Sasaki (DFKI, ERCIM)
Document Location	http://lider-project.eu/?q=doc/deliverables

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	D3.2.1	Title	D3.2.1: Roadmap for the use of linguistic linked data for content analytics – Phase I
Workpackage	Number	WP3	Title	Reference architecture and roadmap

Date of Delivery	Contractual	30.09.2014	Actual	31.10.2014
Status	Final			
Nature	prototype □ report X dissemination □			
Dissemination level	public X consortium □			

Authors (Partner)	Philipp Cimiano, Matthias Hartung, Sebastian Walter (UNIBI) Asunción Gómez Pérez, Guadalupe Aguado de Cea Elena Montiel Ponsoda, Victor Rodríguez Doncel (UPM**)** Dave Lewis (TCD) Paul Buitelaar (NUIG) Roberto Navigli, Tiziano Flati (UNIROMA1)			
Responsible Author	Name	Philipp Cimiano	E-mail	cimiano@cit-ec.uni-bielefeld.de
	Partner	Bielefeld University	Phone	+49 (0) 521 106 12249

Abstract	This documents describes a roadmap for the use of linguistic linked data in content analytics.

(for dissemination)	
Keywords	Roadmap, linked data, content analytics

Version	Modification(s)	Date	Author(s)
01	Final Version	31.10.2014	Philipp Cimiano
02	Post-delivery cleanup + glossary	31.01.2015	Philipp Cimiano

Participants		Contact
Universidad Politécnica de Madrid (UPM, Spain)		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 (TCD, Ireland)		David Lewis Email: dave.lewis@cs.tcd.ie
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)		Felix Sasaki Email: fsasaki@w3.org
National University of Ireland, Galway (NUIG, Ireland)		Paul Buitelaar Email: paul.buitelaar@deri.org
Institut für Angewandte Informatik e.V. (INFAI, Germany)		Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Universität Bielefeld (UNIBI, Germany)		Philipp Cimiano Email: cimiano@cit-ec.uni-bielefeld.de
Universita degli Studi di Roma La Sapienza (UNIROMA1, Italy)		Roberto Navigli Email: navigli@di.uniroma1.it
GEIE ERCIM (ERCIM, France)		Felix Sasaki Email: fsasaki@w3.org

Table of Contents

GLOSSARY

AI	Artificial Intelligence
API	Application Program Interface
BI	Business Intelligence
BPMLOD	Best Practices for Multilingual Linked Open Data
CCG	Combinatorial Categorial Grammar
CEF	Connecting Europe Facility
CRM	Customer Relationship Management
DITA	Darwin Information Typing Architecture
EDF	European Data Forum
GAAP	Generally Accepted Accounting Principles
GDPR	General Data Protection Regulation
GF	Grammatical Framework
IP	Intellectual Property
IPR	Intellectual Property Rights
IR	Information Retrieval
ITS	Internationalization Tag Set
LD	Linked Data
LD4LT	Linked Data for Language Technology
lemon	Lexicon Model for Ontologies
LIDER	Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	Language Resource
LTAG	Lexicalized Tree Adjoining Grammar
MT	Machine Translation
NER	Named Entity Recognition
NIF	NLP Interchange Format

NLP	Natural Language Processing
OWL	Web Ontology Language
POS	Part of speech
QA	Question Answering
R&D	Research and Development
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
Rmp	Roadmap
ROI	Return on investment
RPI	Remote Method Invocation
SKOS	Simple Knowledge Organization System
SME	Small and medium-sized enterprises
SOAP	Service Oriented Architecture and Programming
SPARQL	SPARQL Protocol and RDF Query Language
TBX	TermBase eXchange
UIMA	Unstructured Information Management Architecture
W3C	World Wide Web Consortium
W3C	World Wide Web Consortium
XBRL	eXtensible Business Reporting Language
XML	eXtensible Markup Language

1. Executive Summary

As data is being created at an ever-increasing pace, more and more organizations will exploit insights generated from that data to optimize processes, improve decision making and to modify existing business models or even generate radically new ones.

Data has been, in fact, regarded as the *oil of the new economy*. In this line, data linking and content analytics are the key technologies to refine this oil so that it can drive the motor of many applications. Refinement consists of i) **homogenization**, ii) **linking**, iii) **semantic analysis**, and iv) **repurposing**.

Homogenization makes sure that data is described using agreed terminologies and standardized vocabularies or ontologies in machine-readable formats. Thus, data becomes interoperable and easily exploitable due to the semantic normalization that avoids conceptual mismatches and ambiguities. Refinement also includes **linking** data across datasets and sites (**resulting in linked data**), and is crucial to make data exploitable as a whole rather than as isolated, unrelated datasets. This also includes linking unstructured datasets in different natural languages. In order to exploit data meaningfully, it needs to be **semantically analyzed**. This holds in particular for unstructured data, e.g. textual data from which the key messages and facts need to be extracted and expressed with respect to standardized vocabularies. Structuring and linking unstructured data such as text is referred to as **linguistic linked data** (LLD)**.** Finally, **repurposing** consists of transforming data so that it can be used for a different purpose than it was originally conceived for. Repurposing can include merging and mashing up datasets, format transformations, modifications to the data to fit different audiences (e.g. experts vs. novices), speakers of different languages, etc.

This document presents a roadmap to create the infrastructure that makes all of the above possible. It focuses on three overarching application fields and needs: i) **Global Customer Engagement Use Cases**, ii) **Public Sector and Civil Society Use Cases**, and iii) **Linguistic Linked Data Life Cycle and Linguistic Linked Data Value Chain.**

Regarding **Global Customer Engagement Use Cases**, the challenge for the future will be to create ecosystems in which data from different sources and modalities comes together, and that build the basis for developing omnichannel experiences for customers. This requires linking data across modalities and techniques for repurposing and composing information items into stories and narrations that are more amenable and accessible to users. Consistency of message across channels, languages and audiences will be important and crucially supported by linked data technology. As we observe a shift from marketing activities characterized by a push and active recommendations to a new paradigm of customer engagement that is transparent as it represents a commodity that recognizes and fulfills customer needs in real time, richer linked semantic descriptions of users, products, contexts and intentions will be needed to support matchmaking.

In the area of **Public Sector and Civil Society Use Cases**, linked data can make an important contribution to the creation of a single digital market in which national barriers are overcome and services are interoperable and operate across counties. A crucial ingredient for the creation of a single digital market is the development of ontologies and terminologies that harmonize the concepts used in different countries and jurisdictions, as a basis to reach interoperability and develop a new generation of (public) services that is implemented across countries. This is in spirit to the vision behind the *Connecting Europe Facility (CEF)*. New robust methodologies for alignment of different conceptualizations originating from different cultural, national and linguistic contexts, as well as techniques for collaborative, cross-border ontology engineering will be needed. We also foresee that in the near future we will need a better understanding of the key domains in which cross-lingual and cross-border

communication is urgently needed as a basis to develop shared vocabularies that support language-independent communication in these domains.

Regarding **Linguistic Linked Data Life Cycle and Value Network Requirements**, the future will have to bring an ecosystem in which linguistic resources are easily exploitable by content analytic providers and workflows in a way that provenance, licensing and metadata are clearly exposed to support trustful consumption of resources. Future efforts will need to create principles for a market in which linguistic linked data, both open and closed, can be traded, developing new business models that include also non-monetary transactions. The goal will be to create an ecosystem in which both linguistic resources and services building upon these are i) easily discoverable, ii) trustful and certified, iii) comparable and benchmarkable, iv) easily composable and exchangeable, v) multilingual, and vi) scalable. This involves two challenges: i) bringing the stakeholders together to establish principles for such a market, and ii) develop a technical infrastructure including standardization of APIs and vocabularies that supports plug and play principles.

The first draft of this document has been created by LIDER project partners (FP7 CSA, refererence number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies). The analysis builds on the findings and forecasts of a number of existing public reports on the topic, by aggregating and analyzing different needs and forecasts expressed in these documents as a basis to identify application areas in content and big data analytics where linked data could represent a key enabling technology. Building on the insights about application areas where the potential of linked data technology, in particular linguistic linked data technology, is regarded as very high, a roadmap is defined that extrapolates the above mentioned findings to define an R&D roadmap that can support research organizations, enterprises and funding agencies in decision making and to prioritize R&D investments. A further goal for this roadmap is to define an R&D agenda at the intersection of the following communities: language resources, natural language processing, and Big Data.

2. Background and Context

This roadmap is a product of four roadmapping events organized by the LIDER project (see also https://www.w3.org/community/ld4lt/wiki/Lider_roadmapping_activities):

1. 1st LIDER Roadmapping Workshop in Athens, March 21, 2014, collocated with the European Data Forum (EDF), **with 43 participants**
2. 2nd LIDER Roadmapping Workshop in Madrid, May 8-9, 2014, collocated with the Multilingual Web Workshop (MLW), **with 44 participants**
3. 3rd LIDER Roadmapping Workshop in Dublin, June 14, 2014, collocated with Localization World Dublin, **with 40 participants**
4. 4th LIDER Roadmapping Workshop in Leipzig, 2nd of September, 2014, collocated with SEMANTICS, **with 51 participants**

Further, the LIDER project has interacted with relevant stakeholders in the context of a number of community groups including:

1. W3C Community Group on Linked Data and Language Technologies (LD4LT), with **79 participants**
2. W3C Community Group on Best Practices for Multilingual Linked Open Data (BMLOD), with **79 participants**
3. W3C Community Group on Ontology Lexica (Ontolex), with **90 participants**

Thus, this roadmap is based on the needs, use cases and predictions of **hundreds of relevant stakeholders** from both academia and industry, which have been collected through the above mentioned events and community groups.

This document represents a snapshot of the Living Roadmap, to which parties active in the LD4LT community group can contribute.

3. Introduction

Content is growing at an impressive, exponential rate. Exabytes of new data are created every single day (Pepper et al. 2014). In fact, data has been recently referred to as the "*oil"* of the new economy (Palmer et al. 2006), where the new economy is understood as *"a new way of organizing and managing economic activity based on the new opportunities that the Internet provided for businesses"* (Alexander 1983).

There are several indicators that clearly corroborate that the exponential growth of data will continue:

- **Volume**: The data streams already generated today are huge. Only one hour of customer transaction data at Wal-Mart, corresponding to 2.5 petabytes, provides 167 times the amount of data housed for example by the Library of Congress (Bilbao-Osorio et al. 2014).
- **Growth Rate**: 90% of the data available today has been generated in the past two years only (SINTEF, 2014). The International Data Corporation (IDC) estimates that all digital data created, replicated or consumed will grow by a factor of 30 between 2005 and 2020, doubling every two years. By 2020, it is assumed that there will be over 40 trillion gigabytes of digital data, corresponding to 5,200 gigabytes per person on earth (Gantz and Reinsel 2012).
- **Internet of Things**: Cisco estimates that currently less than 1% of physical objects are connected to IP networks. However, this is estimated to change radically to up to 50 billion devices connected to the Internet by 2020, corresponding to between 6 and 7 devices per person on the planet (Cisco 2013). These 50 billion devices will constantly generate data at a scale without precedent.

Content analytics, i.e. the ability to process and generate insights from existing content, plays and will continue to play a crucial role for enterprises and organizations that seek to generate value from data, e.g. in order to inform decision and policy making.

A basic distinction can be made between **structured and unstructured** data. Structured data is data that follows a given pre-defined schema or data model, such as data in standard relational or non-relational (including NoSQL) databases, or data expressed in Web languages such as the Resource Description Framework (RDF). Unstructured data does not follow a predefined schema and comprises texts, blogs, pictures, and sensor data.

Current estimates suggest that only **half a percent of all data is being analyzed to generate insights** (Gantz and Reinsel 2012). Furthermore, the **vast majority of existing data is unstructured and machine-generated** (Canalys 2012), with data automatically generated by mobile devices and sensors constituting the majority.

As corroborated by many analysts, substantial investments in technology, partnerships and research are required to reach an ecosystem consisting of many players and technological solutions that provide the necessary infrastructure, expertise and human resources required to make sure that organizations can effectively deploy content analytics solutions at large scale in order to generate relevant insights that support policy and decision making, or even to define completely new business models in a data-driven economy.

Assuming that such investments need to be and will be made, **this report explores the role that linked data and semantic technologies can and will play in the field of content analytics and will generate a set of recommendations** for organizations, funders and

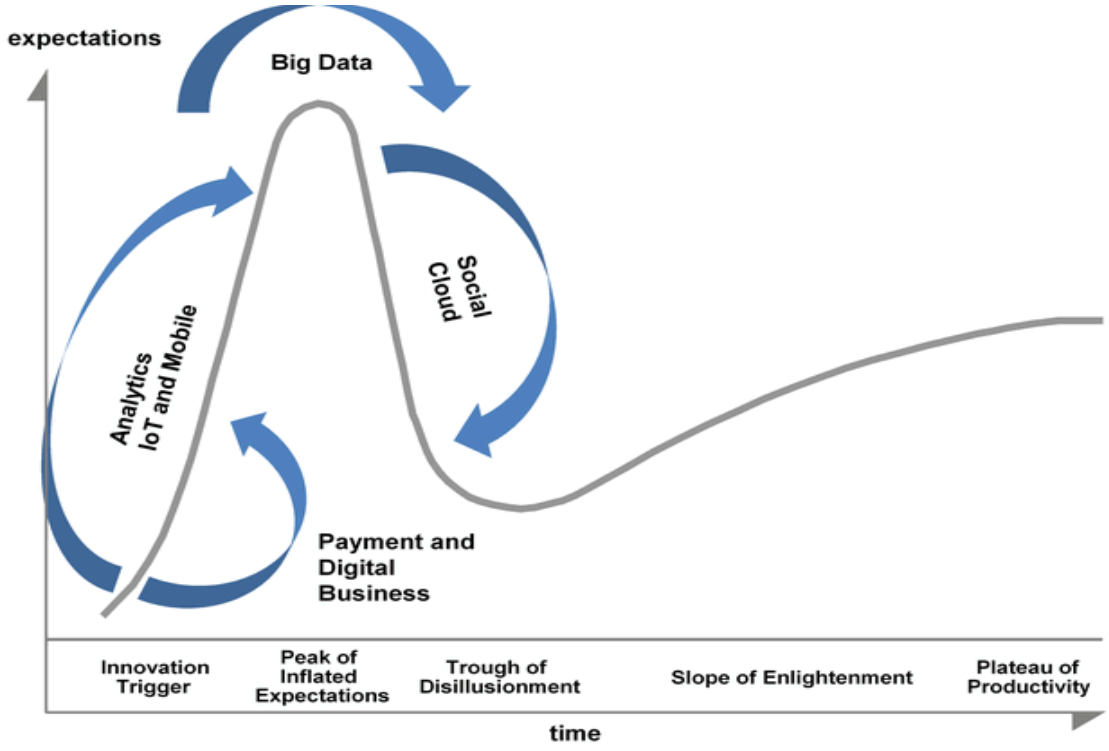
researchers on which technologies to invest as a basis to prioritize their investment in R&D as well as on optimizing their mid- and long-term strategies and roadmaps.

The main sources this report draws upon are the following ones:

- Forrester Research Report: TechRadarTM for AD&D Pros: Digital Customer Experience Technologies, QS 2013 (Yakkundi et al. 2013)
- iVIEW Report by the International Data Corporation (IDC): The Digital Universe in 2020: Biig Data, Bigger Digital Shadows, and Biggest Growth in the Far East (Gantz and Reinsel 2012)
- The Global Information Technology Report 2014, published by the World Economic Forum (Bilbao-Osorio et al. 2014)
- Gartner Hype Cycle Special Report 2014
- Text Analytics 2014: User Perspectives on Solutions and Providers by Alta Plana Corporation (Grimes 2014)
- Report Rethinking Personal Data: A New lens for Strengthening Trust, published by the World Economic Forum (2014)
- The LT-Innovate Innovation Manifesto, published by LT Innovate (2014)
- Strategic Research Agenda for Multilingual Europe 2020, published by the META Technology Council (2012)
- Results of the Roadmapping Workshops organized by the LIDER project
- IEEE CS 2022 Report
- Call for Papers from the linked data and Semantic Web communities

The report is structured as follows: In Section 4 we discuss general IT trends as identified by Gartner in order to position this roadmap with respect to major trends in the field of IT. Section 5 summarizes the current needs in content analytics, focusing on the survey carried out by Alta Plana (see Section 5.1). Further, we analyze the needs of the application development and delivery industry as identified by Forrester's TechRadar^{TM} report (Yakkundi et al. 2013) in Section 5.2. Finally, we summarize the main outcomes of the 4th LIDER roadmapping workshop organized by the LIDER consortium as part of the MLODE workshop and collocated with the SEMANTICS conference (see Section 5.3). Section 6 briefly discusses the vision and objectives behind the EC's Connecting Europe Facility (CEF) program and how linked data technology can contribute to this vision. The actual roadmap is presented in Section 8; it mentions the most promising application areas and future directions for the role of linked data, and in particular linguistic linked data, in content analytics. Section 9 concludes and discusses ways forward, providing some recommendations for funders and researchers.

4. General IT Trends



As a basis to identify general IT trends, we consider Gartner's well-known Hype Cycle from 2014. A visualization of the cycle can be found in the figure above. The hype cycle focuses on newly emerging technologies as they move into mainstream adoption.

The topics mentioned in the Hype Cycle of 2014 are the following seven technologies:

- **Mobile**: According to the analysis of Gartner, mobile technology will become the main vehicle for business applications, allowing organizations to reach more users than via other conventional channels. Due to the pervasiveness of mobile devices among customers, mobile technology is creating disruptive opportunities for business. Thus, Gartner is placing mobile technology as a technology moving rapidly to the peak.
- **Internet of Things (IoT)**: The Internet of Things (IoT) is also considered as a technology moving rapidly to the peak. Given that many companies are defining a strategic agenda for digital business, it is only logical that technologies operating on the physical world will go digital and become part of the network. This is expected to have a large impact and even be transformational with respect to digital business models and production processes, e.g. in the manufacturing area (Industry 4.0).
- **Analytics**: As more and more data is generated, e.g. through devices connected to the network as part of the Internet of Things (see above), analytics over these data streams will be an essential ingredient. Analytics technologies are thus regarded as moving into the peak. Gartner foresees that the delivery of analytics as a service (business analytics PaaS) will be a major and important trend. A further trend can be observed in the convergence of information technology (IT) and operational technology (OT), corresponding to the growing use of IT solutions in OT vendors' products. The primary driver in bringing the two worlds together is the need to use analytics from diverse data to improve decision making across the supply chain.
- **Big Data**: Gartner predicts that Big Data is moving over the peak. Big Data is related to cost-effective information processing of high-volume, high-velocity and highly varied information assets to generate enhanced insights and thus support decision making. While the interest and demand for Big Data solutions is still undiminished, according to Gartner it is moving over the peak due to the convergence to a set of

promising solutions and approaches. The movement of Big Data topics over the peak and the movement of analytics towards the peak clearly suggest that the hype and interest in information and data processing is fostering adoption of these technologies in the value chain.

- **Cloud**: According to Gartner, cloud technology is generally moving towards the trough. Key topics in the area of cloud technology which are still climbing the peak are *cloud access security*, *cloud application development services*, and *cloud integration platform services*. Topics that are moving over the peak towards the trough include c*loud computing, private cloud computing, and hybrid cloud* (referring to the coordinated use of cloud services across isolation and provider boundaries among public, private and community service providers, or between internal and external cloud services). *Mobile cloud*, referring to the use of cloud services for information sharing across devices, has moved quickly over the peak to the trough, indicating maturity for mainstream adoption. *Personal cloud services*, defined as the individual's collection of digital content, are moving away from the peak, due to their main usage for storage and synchronization without providing further significant value.
- **Social**: Social technologies are regarded as having shown a significant shift from peak position to the trough. This is mainly due to the fact that users and vendors have realized that generating business value on the basis of social technologies is more challenging than expected. The movement to the trough, on the other hand, clearly indicates that some social technologies are becoming mainstream, in particular in the area of digital business. Nevertheless, it is foreseen that social technologies, services and disciplines will soon become mainstream technologies within digital businesses. A technology that is expected to move towards the plateau are *peer-to-peer communities*, i.e. virtual collaboration environments fostering collaboration among people and organizations outside the enterprise. Technology for analyzing social networks has moved into the trough due to the difficulty of collecting relevant and reliable data and because turning knowledge into action has been found to be difficult. Some technologies have completely fallen down toward the trough, including s*ocial gaming*, *social TV* and *social profiles*.
- **Payment and digital business**: Technologies for integrating payment methods with loyalty programs have moved off the trigger position to midpoint position. Digital Business profiles are placed by Gartner right below the peak of inflated expectations, due to the high number of organizations claiming to have a digital strategy aligned with their overall business strategy.

4.1 Summary

The fact that the topics *social* and *cloud* are identified by Gartner as moving over the trough to the plateau of maturity clearly shows that these technologies are maturing and becoming part of the IT value chain. Gartner expects that by 2016 *cloud* and *social* will be so pervasive that 60% of organizations will have adopted them. The impact that *cloud* and *social* have on enabling organizations to run their operations more efficiently, drive new capabilities, serve customers and partners effectively, and respond to disruptive threats and opportunities in the market more rapidly is expected to continue to be key for organizations. The focus in the next years will be on understanding how social and cloud-based solutions can be applied to generate added value. Any roadmap such as this one in the area of IT will have to consider the prominent role that these two technologies will continue to play in the future. The third relevant technology, Big Data, is regarded by Gartner as moving beyond the peak of inflated expectations, leading to early signs of technological convergence and adoption. Big data can surely be regarded as a key trend in IT that will soon lead to mature solutions that generate valuable insights for enterprises that understand how to effectively make use of these techniques. In this line, IBM has predicted that most companies will have a dedicated role of a data scientist that oversees the organization and understands how to exploit data to optimize the business (Teerlink et al. 2014).

5. Needs in Content Analytics

This section analyses and summarizes the current main needs in content analytics industries and applications. It summarizes the main results from a survey carried out by Seth Grimes from Alta Plana corporation in 2014 in order to understand current needs in content analytics. We further summarize the main challenges identified by Forrester in their *TechRadar*™ report (Yakkundi et al. 2013). Finally, we reflect on the main results from the 4th LIDER roadmapping workshop that was organized by the LIDER consortium in the context of the MLODE workshop in Leipzig, which, in contrast to other roadmapping workshops organized by LIDER, had a strong focus on content analytics.

5.1 Survey of Text Analytics Needs

This section analyzes the top needs in the field of content analytics, focusing in particular on text analytics. The summary of needs here is based on the report "*Text Analytics 2014: User Perspectives on Solutions and Providers"* by Seth Grimes from Alta Plana (Grimes 2014). The survey described in the above report gathers responses from 200 participants from the content analytics sector. The survey targeted or prospect users of text analytics technologies, integrators, consultants and managers as well as executive staff in those roles. Researchers and developers applying text analytics technologies were also invited to participate.

The four technology-related growth drivers for text analytic solutions identified by the survey are the following ones:

- **Open source:** Open source technology is expected to lower barriers both to technology adoption for researchers and more sophisticated users.
- **API economy:** The availability of hosted, on-demand API-based web services is expected to lower entry barriers and provides enormous flexibility for adopters.
- **Data availability:** The availability of data is crucial and increasing.
- **Synthesis:** Providing support for summarizing and synthesizing datasets, and providing answers to specific information needs is crucial.

The **five key market drivers**, i.e. application fields, that have the potential to generate business value identified in the survey are the following ones:

- **Customer interactions,** including support for customer services and customer experience by deploying text analytics to transitional channels such as contact centers but increasingly also to social media, with the goal of optimizing service and fostering engagement.
- **Omnichannel solutions,** comprising the analysis and aggregation of data from different channels, e.g. survey, social media, news warranty, chat, contact center data, etc.
- **Consumer and market insights,** consisting of the deployment of text analytics solutions for market research and the generation of insights about market and customer needs, trends, and opinions. An observable trend is that social data is regarded as increasingly reliable and as a trusted source that can complement classical survey data and deliver complementary insights.
- **Search and search-based applications,** providing new search functionalities that go beyond enterprise search or online information retrieval to provide a platform for high-value applications that includes advertising, e-discovery and compliance, business intelligence and customer self-service.
- **Health care and clinical medicine,** providing new analytics solutions, e.g. supporting diagnosis or analysis of claims. These new solutions are expected to drive the market and go beyond mere text mining of scientific literature.

The key data sources relevant for text analytics solutions are:

- blogs and other social media (mentioned by 61% of respondents)

- news articles (42%)
- comments on blogs and articles (38%)
- online forums (36%)

Clearly, user generated content ranks highest and can be assumed to continue to play a key role in the near future.

The top business applications identified are the following ones (considering only those mentioned by at least 20% of the respondents):

- voice of the customer (mentioned by 39% of respondents)
- brand/product/reputation management (38%)
- competitive intelligence (33%)
- search, information access or question answering (29%)
- Customer Relationship Management (CRM) (27%)
- Content management or publishing (25%)
- Online commerce (16%)
- Life sciences or clinical medicine (15%)
- E-discovery (14%)
- Insurance, risk management, or fraud detection (13%)

Summarizing the written comments provided by the respondents to the survey, we can identify the following bottlenecks and **limitations of current content analytics technology**:

- **Customization:** A major issue is the required effort to customize existing solutions to a particular domain and application. A lack of domain-specific models has been identified.
- **Usability**: Usability of current text analytics solutions is generally regarded as low, with APIs that are difficult to use. Deployment of text analytics requires in-house expertise on natural language processing and data mining, which is clearly a bottleneck for adoption.
- **Limitations**: Accuracy of current solutions as well as depth of analysis is quite limited (see also Cieliebak et al. 2013).
- **Cost/Effort**: The cost and effort needed in selecting appropriate vendors and solutions, integrating their solutions, analyzing results and generating valuable insights to obtain a return on investment (ROI) requires huge effort and cost. The learning curve for deployment of content analytics solutions is generally too steep.
- **Integration**: Integration into existing workflows or systems, in particular existing Business Intelligence solutions, is regarded as difficult and requires a major effort.

The following information types are highly relevant in content analytics:

- topics and themes (mentioned by 66% of respondents)
- sentiments, opinions, attitudes, emotions, perceptions, intent (54%)
- relationships and/or facts (47%)
- named entities (56%)
- concepts (51%)
- document metadata (47%)
- other entities (34%)
- events (33%)
- semantic annotation (31%)

Important properties of content analytics solutions are:

- ability to generate categories or taxonomies (65% of respondents)
- ability to use specialized dictionaries, taxonomies, ontologies or extraction rules (54%)

- broad information extraction capability (53%)
- document classification (53%)
- deep sentiment, emotion, opinion, intent extraction (45%)
- low cost (44%)
- real-time capabilities (43%)
- sentiment scoring (41%)
- support for multiple languages (40%)
- open source (37%)
- predictive analytics integration (36%)
- big data capabilities (33%)
- ability to create custom workflows (33%)
- Business Intelligence integration (32%)
- sector adaptation (30%)
- support for data fusion (28%)
- hosted or Web service (on-demand API) (25%)
- media monitoring/analysis interface (22%)

Participants also mentioned the need to process content in languages other than English. The top languages (other than English) mentioned by at least 10% of respondents are:

- Spanish (mentioned by 38% of respondents)
- French (36%)
- German (34%)
- Italian (18%)
- Chinese (16%)
- Portuguese (13%)
- Arabic (10%)

5.1.1 Summary

The main applications for textual content analytics are in the areas of i) **analyzing the voice of the customer** in the context of Customer Relationship Management (CRM), ii) **brand, product and reputation management**, iii) **technology surveying and competitive intelligence**, iv) **content management and publishing**, and v) **search, information access and question answering**. The main source of information is clearly user-generated content (blogs, social media, online forums etc.). While the main tasks required are rather conventional (topic extraction, document classification, entity extraction, relation, event extraction etc.), there is a clear need for analytics solutions that i) **are tuned to the needs of particular domains**, and ii) **are able to generate and incorporate semantic domain-specific knowledge** in the form of taxonomies, terminologies etc. to support domain customization.
The main issue with current text analytics technologies identified are i) **a lack of standard and flexible APIs**, ii) **the high effort and resources needed for customization and domain adaptation**, iii) **the level of expertise required to deploy and integrate solutions into own workflows**, and iv) **the lack of accuracy and depth in analysis** (see also Cieliebak et al. 2013). Multilinguality is a further very important topic. Language coverage needs to be extended to address the most needed languages: Spanish, French, German, Italian, Chinese, Portuguese and Arabic. Future generations of text analytics solutions will thus have to provide easy-to-use standard and flexible APIs, increase accuracy and depth of analysis, and be able to process the languages mentioned above. The ability to integrate additional background knowledge to make domain adaption easier will also be a key capability. Finally, solutions need to be easily integrateable into existing workflows, processes and tool chains.

5.2 Application Development and Delivery (AD&D)

This section analyzes the main findings of the TechRadar™ report by Forrester (Yakkundi et al. 2013). Many digital companies are concerned with the challenge to invest and maintain an

ecosystem of technology that supports digital customer experiences that are enjoyable, innovative and contextualized. Forrester defines the sum of these digital customer experience technologies as *"a solution that enables the management, delivery and measurement of **dynamic, targeted, consistent** content, offers products, and service interaction **across digitally enabled customer touch points**"* (emphasis is ours).

The relevant technologies identified in the TechRadar™ report are the following ones:

- **A/B and multivariate testing**: Multivariate testing allows marketers and digital experience professionals to test different components of a digital experience, such as site design, site usability, campaign landing page with respect to some measure, and the like.
- **Digital asset management**: Digital asset management is software used to manage the creation, production, management, distribution and retention of rich media content including audio, video, graphical images and compound documents.
- **eCommerce**: eCommerce solutions provide companies with the capability to connect with the market, sell to and serve B2B and B2C customers across many digital touch points.
- **Email marketing platforms**: These tools help to build, execute and monitor email advertising campaigns.
- **Mobile analytics**: These analytics tools support the collection, analysis and measurement of mobile app traffic and user data to optimize mobile experience.
- **Online video platforms**: These are solutions that focus on the distribution of video content.
- **Optimization**: Optimization refers to a set of tools that leverage exploratory, descriptive and predictive statistical techniques to drive relevant content, interactions with and offerings to end users.
- **Portals**: Portals are software platforms that aggregate content, data and applications, and deliver them in a rich, personalized environment that supports digital customer experience.
- **Product content management**: Product content management allows for identifying or deriving trusted product data and content across heterogeneous data environments.
- **Recommendation engines**: These tools help to recommend products, texts or other content to visitors on different types of websites with the goal of delivering the most relevant and useful experience to the customer.
- **Site search**: Site search refers to tools that offer search capabilities over content from different systems in a portal or website.
- **Social depth platforms**: Tools that integrate social content and experiences into marketing sites.
- **Web analytics**: Tools supporting the collection of usage data in Web channels.
- **Web content management**: Web content management software supports organizations in creating websites and online experiences as well as in creating, managing and publishing digital content across websites or multiple channels.

The holy grail of effectiveness identified by Forrester is to create contextualized, personalized multi-channel experiences for customers. The following key success factors can be identified:

- **Multi-channel experience**: Multiple channels have to be targeted in order to deliver a consistent message, possibly in different languages and modalities, integrating Web content management with product management and eCommerce solutions.
- **Contextualization**: Portals should be extended to manage and deliver relevant, tailored experiences including personal or individual information on a per-user basis.
- **Holistic recommenders**: Rather than only performing product recommendations, it will be crucial to create recommender systems that provide pervasive support for users.
- **Breaking down silos**: Customer data, product data and social media data needs to be brought together to deliver contextual cross-channel experiences.

- **Adaptability and control**: Marketing experts and line-of-business people need more control over digital channels in order to react dynamically to arising customer needs.

5.2.1 Summary

A key challenge in the area of application development and delivery consists in **the combination of multiple channels to effectively target customers across channels**. On the one hand, this requires to make sure that **consistent messages are delivered across channels**, which can be ensured by incorporating terminological and lexical resources for the particular domain and application. A further challenge is to foster the **convergence of different subsystems** (e.g. content management, product management, and customer management) into a rich and **seamless semantic ecosystem that can support rich customer experiences and contextualized, personalized and situated media delivery and recommendations**. A further key challenge consists in **supporting agility in the capture, integration and provision of content**, allowing marketing experts to dynamically react to changing customer and market needs.

5.3 4th LIDER Roadmapping Workshop

The LIDER project organized a roadmapping workshop in Leipzig on September 2nd, 2014. This workshop was part of the Multilingual Linked Open Data for Enterprises (MLODE) workshop and collocated with the SEMANTICS conference. The program of the roadmapping workshop can be found online at `http://mlode2014.nlp2rdf.org/lider-roadmapping-workshop/`. The workshop featured the contributions of 12 companies and users from the area of content analytics, which shared their view on current challenges and perspectives for the exploitation of linked data in content analytics tasks.

The main themes identified at the roadmapping workshop were the following five (prioritized with respect to importance according to the vote of the audience at the workshop): i) **resource creation and sharing**, ii) **open linked publishing and consumption**, iii) **multilingual semantic content analytics and search**, iv) **standardization of APIs**, and v) **big text and data analytics**.

5.3.1 Resource Creation and Sharing

One of the recurring topics at the roadmapping workshop was the need for and lack of appropriate linguistic resources to facilitate training of content analytics solutions in order to adapt them to specific domains. Industry participants stressed that business-friendly licenses that allow for the exploitation of data in commercial contexts are urgently needed. Resources that were most frequently mentioned as important were terminologies, lexicographic resources, POS-tagged datasets (especially for user-generated content, e.g. tweets), treebanks, datasets annotated with sentiment, NER-annotated corpora as well as corpora in which named entities are linked to external knowledge bases or resources. In general, it was observed that while for most applications resources of mid-level quality are sufficient, for some applications in which quality is crucial, highly curated and verified resources are needed. Lexica are a good example for this, ranging from largely automatically created resources (e.g. BabelNet or automatically generated WordNets) to highly curated and manually validated lexical resources such as Princeton WordNet and KDictionaries. An important question is how new paradigms and approaches to resource creation can find an optimum in the quality vs. cost tradeoff and how human-machine collaborative workflows can be defined that increase the quality of resources while at the same time minimizing the amount of work needed by experts. The community of content analytics solution vendors clearly mentioned the need for relevant resources for micro-domains, i.e. the availability of highly domain-specific resources (lexica, terminologies, annotated corpora, ontology of intentions) that would support domain adaptation and that could be widely reused across vendors.

5.3.2 Open Linked Data Publishing and Consumption

With respect to publishing data as linked open data, a nuanced perspective emerged during the roadmapping workshop. While having high-quality **open and reusable data** was clearly seen as a benefit (see the idea of an *open commons* mentioned by Alex Pentland (Pentland 2014)), it was also mentioned that the publication of linked data should take into account the principle of economy and efficiency, in particular taking into account that the cost for publishing, quality control and the like should not exceed the added value provided by the resource. Further, many internal datasets at organizations or companies can not be published because doing so would disclose private information about employees or customers. In addition, many datasets are specific to the particular structures and processes implemented in a certain organization and, when taken out of this context, do not provide any added value. In the future, more experiences are needed to understand the cost and value of publishing a certain dataset as linked data as well as metrics to monitor and quantify its impact and reuse. Methodologies that simplify the process of publishing data as linked data are also needed. In general, it was mentioned that the linked data effort should focus more on *linking* rather than on publishing isolated datasets on the Web. In general, it was also mentioned that instead of focusing on the publication of large multi-purpose datasets, the focus in the future should be on publishing small and reusable buildings blocks that can be used for a specific but frequently reoccurring purpose. An important target would be to **link concepts across languages**, effectively creating a multilingual linked knowledge infrastructure that can be exploited in applications that need to provide support across languages.

5.3.3 Multilingual Semantic Content Analytics and Search

The need for robust and accurate text analytics solutions is still pressing. As identified by the MetaNet activities (see `http://www.meta-net.eu/whitepapers/press-release`), the support for most European languages in terms of linguistic resources and natural language processing tools is low. This situation was confirmed by the participants of the roadmapping workshop. Robust and accurate text analytics solutions are needed at different levels: part-of-speech tagging, chunking, parsing, named entity recognition, named entity linking, information extraction, sentiment analysis, and so on. In general, participants expressed the clear need to go deeper and have text analytic solutions that extract deeper semantics including pragmatics, e.g. to identify the intention of a customer. Solutions that can perform cross-lingual normalization of content, terms, and named entities are urgently needed, where translation is only one component of the needed technology infrastructure. Semantic search (e.g. via keywords or question answering) at the meaning level rather than at the string level is still one of the major needs of the market, as mentioned by several participants of the roadmapping workshop.

5.3.4 The Human Factor

The human factor was identified as an important aspect for the adoption and proliferation of content analytics solutions. In many cases, **the expectations of potential end users lead to disappointment with the current state of the art of content analytics solutions**. Instead of focusing on reporting performance figures of single tools, **experiences in which performance figures are needed for which type of application need to be gathered**, and end users need to be sensitized that tools are far from perfect but that the performance levels can be sufficient for a particular purpose or application. In general, it was felt that the community of content analytics vendors, developers and researchers need to invest in raising awareness of the benefits, limitations, success stories but also pitfalls of content analytics solutions, and devise effective ways of communicating these aspects to potential customers and users.

5.3.5 Standardization of APIs

An important bottleneck in the field of content analytics is that all providers offer their own proprietary API, thus hindering the easy exchange of solutions and the composition of different solutions into more complex workflows. This makes the comparison of different solutions very difficult and further leads to *vendor locking*, as once a company has adopted a certain provider, given the high costs of adaptation (see Section 5.1), it is likely to stay with that vendor. The participants of the workshop expressed the desideratum of working towards **standardization of APIs in content analytics**. For one thing, this would make it easier to **integrate different solutions into one complex product** and would also **support joint partnerships between different providers of content analytics solutions**. Second, from the point of view of the customer, it would **support cross-vendor comparison of technology to make more informed decisions**. Finally, standardized APIs would also **facilitate benchmarking and quality assessment of tools and services** by running automated tests or evaluations, both by the research community and as part of an open ecosystem of resources and tools.

5.3.6 Big Text and Data Analytics

Surprisingly, the topic of Big Data analytics was prioritized lowest by the participants at the roadmapping workshop. One explanation for this is that Big Data as a hype has generated already some disappointment (this is in line with Gartner predicting Big Data to move down from the peak of inflated expectations, see Section 4). The need to process large amounts of data is clearly there and likely to move into mainstream and adoption, but the hype seems to be decreasing. The challenge of developing solutions that can scale to large streams of unstructured data is still a crucial one. Especially the integration of knowledge across languages and formats at a larger scale is an important challenge to addresss

5.3.7 Summary

The 4th LIDER roadmapping workshop was organized in Leipzig on September 2nd as part of the MLODE workshop and collocated with the SEMANTICS conference. Several companies and users from the content analytics area joined the roadmapping workshop and active discussions happened among and with the participants in order to face current challenges for the exploitation of linked data in content analytics tasks.
Six main themes were identified during the workshop, ranging from resource creation and sharing (where participants pointed out the lack and need of appropriate linguistic resources for content analytics solutions) to linked open data publishing (with an outcry for economy and efficiency of the publication of linked data), as well as the need for support of resource-poor European languages and the need to raise awareness about the benefits and limitations of content analytics solutions among customers and adopters.

6. Connecting Europe Facility

The goal of the Connecting Europe Facility (CEF) program by the European Commission is to support the development of *"high-performing, sustainable and efficiently interconnected trans-European networks in the fields of transport, energy and digital services"* (European Commission, 2012). By means of this, the European Commission expects to contribute to increase growth, jobs and competitiveness for Europe. A budget of 50 billion EUR between 2014 and 2020 is foreseen.

The stated strategic objective of CEF is to contribute to the ***development of a single digital market*** and to effectively **eliminate market fragmentation, making sure that cross-border public services are broadly available and accessible** by millions of citizens and companies to connect to a single market.

A further goal is to contribute to the transformation of Europe into a knowledge-intensive, low-carbon and highly competitive economy. CEF is thus investing in the creation of modern and flexible energy, transport and digital infrastructure networks. The preferred instrument for this are public-private partnerships funded through innovative financial instruments that make investment in infrastructure projects attractive.

With respect to the development of digital service infrastructures, the CEF Digital Service Infrastructures are expected to act as platforms on which innovative applications can be created and deployed, and to facilitate mobility of citizens working across borders. **To overcome service fragmentation and lack of interoperability due to national borders**, the goal is to develop **pan-European services that interoperate across borders and defragment the market**, in particular in the areas of eGovernment, eProcurement and eHealth.

Funding via Horizon 2020 and CEF can contribute to reducing fragmentation between content analytics solutions. By implementing the needs for content analytics described in Section 5, it can help to make data and services operable across national borders. Linked data technologies, in particular linguistic and possibly multilingual linked data technologies, can be expected to play a major role. They can contribute to the interoperability of services, e.g., by providing a means to align different conceptualizations or ontologies. This will **improve data exchange and semantic interoperability**.
For more information on the CEF digital agenda, see `http://ec.europa.eu/digital-agenda`.

7. Linked Data in Research

In order to identify relevant topics that are on the current research agenda, we have analyzed the calls for papers of the year 2014 of four major conferences in the fields of semantic and linked data technologies: The World Wide Web Conference (WWW 2014), which featured a dedicated Semantic Web track, the International Semantic Web Conference (ISWC 2014), the European Semantic Web Conference (ESWC 2014), and SEMANTICS 2014.

Besides looking at the major conferences, we have also identified a number of workshops specifically dedicated to linked data issues:

- Linked Data on the Web Workshop (LDOW2014), collocated with WWW 2014
- Workshop on Linked Data in Linguistics (LDL-2014), collocated with LREC 2014
- Workshop on Semantic Web Enterprise Adoption and Best Practice (WASABI)
- Workshop on Linked Data Quality, collocated with SEMANTICS 2014
- 1st Workshop on Linked Data for Knowledge Discovery, collocated with ECML/PKDD 2014
- Workshop on Linked Open Data 2014 : Improving SME Competitiveness and Generating New Value, collocated with SEMANTICS 2014

7.1 World Wide Web Conference (WWW 2014)

The relevant topics for WWW 2014 in the Semantic Web track were the following ones:

- Infrastructure: Storing, querying, searching, serving Semantic Web data
- Linking, joining, integrating, aligning/reconciling Semantic Web data and ontologies from different sources
- Tools for annotation, visualization, interacting with Semantic Web data, building ontologies
- Knowledge representation: Ontologies, representation languages, reasoning on the Semantic Web

- Applications that produce or consume Semantic Web data, including those in enterprises, education, science, medicine, mobile, web search, social networks, etc.
- Extracting Semantic Web data from web pages and other sources
- Methodologies for the engineering of Semantic Web applications, including uses of Semantic Web formats and data in the development process itself

7.2 European Semantic Web Conference (ESWC 2014)

ESWC 2014 included the following topics in the Linked Open Data Track (emphasis is our own):

- Linked open data extraction and publication
- Storage, publication and validation of data, links, and embedded linked open data
- Linked data integration/fusion/consolidation
- Database, IR, NLP and AI technologies for linked open data
- Creation and management of linked open data vocabularies
- Linked open data consumption
- **Linked data applications (e.g., eGovernment, eEnvironment, or eHealth)**
- **Dataset description and discovery**
- Searching, querying, and reasoning in linked open data
- Analyzing, mining and visualization of linked open data
- Usage of and social interactions with linked open data
- Dynamics of linked open data
- Architecture and infrastructure
- **Provenance, privacy, and rights management; relationship between linked open data and linked closed data**
- **Assessing data quality and data trustworthiness**
- Scalability issues of linked open data

7.3 International Semantic Web Conference (ISWC 2014)

The call for papers for ISWC 2014 included the following topics (emphasis is our own):

- Management of Semantic Web data and linked data
- Languages, tools, and methodologies for representing and managing Semantic Web data
- Database, IR, NLP and AI technologies for the Semantic Web
- Search, query, integration, and analysis on the Semantic Web
- Robust and scalable knowledge management and reasoning on the Web
- **Cleaning, assurance, and provenance of Semantic Web data, services, and processes**
- **Information Extraction from unstructured data**
- **Supporting multilinguality in the Semantic Web**
- **User Interfaces and interacting with Semantic Web data and linked data**
- Geospatial Semantic Web
- Semantic sensor networks
- Query and inference over data streams
- Ontology-based data access
- Semantic technologies for mobile platforms
- Ontology engineering and ontology patterns for the Semantic Web
- Ontology modularity, mapping, merging, and alignment
- Social networks and processes on the Semantic Web
- **Representing and reasoning about trust, privacy, and security**
- Information visualization of Semantic Web data and linked data
- **Personalized access to Semantic Web data and applications of Semantic Web technologies**
- Semantic Web and linked data for cloud environments

7.4 Linked Data on the Web Workshop, collocated with WWW 2014

The Linked Data on the Web Workshop (LDOW), collocated with the World Wide Web Conference 2014, called for the following topics (emphasis is our own):

- Mining the Web of linked data
  - Large-scale derivation of implicit knowledge from the Web of linked data
  - Using the Web of linked data as background knowledge in data mining
  - Integrating large numbers of linked data sources
  - Linking algorithms and heuristics, identity resolution
  - Schema matching and clustering
  - Data fusion
  - Evaluation of linking, schema matching and data fusion methods
- **Quality evaluation, provenance tracking and licensing**
  - **Evaluating quality and trustworthiness of linked data**
  - Profiling and change tracking of linked data sources
  - **Tracking provenance and usage of linked data**
  - **Licensing issues in linked data publishing**
- **Linked data publishing, authoring and consumption**
  - Mapping and publication of various data sources as linked data
  - Authoring and curation of linked data
- **Linked data consumption interfaces and interaction paradigms**
- Visualization and exploration of linked data
- **Linked data applications and business models**
  - Application showcases including browsers and search engines
  - **Marketplaces, aggregators and indexes for linked data**
  - **Business models for linked data publishing and consumption**
- Linked data as pay-as-you-go data integration technology within corporate contexts
- Linked data applications for life-sciences, digital humanities, social sciences etc.

7.5 Workshop on Linked Data in Linguistics (LDL-2014)

The Workshop on Linked Data in Linguistics (LDL-2014), collocated with LREC 2014, mentioned the following topics in their call for papers (emphasis is our own):

- **Use cases and project proposals for the creation, maintenance and publication of linguistic data collections that are linked with other resources**
- Modelling linguistic data and metadata with OWL and/or RDF
- Ontologies for linguistic data and metadata collections
- Applications of such data, other ontologies or linked data from any subdiscipline of linguistics
- Descriptions of data sets, ideally following linked data principles
- **Legal and social aspects of linguistic linked open data**

7.6 Workshop on Semantic Web Enterprise Adoption and Best Practice (WASABI)

The Workshop on Semantic Web Enterprise Adoption and Best Practice (WASABI), collocated with EKAW 2014, had a special focus on Linked Data Lifecycle Management and called for the following topics (emphasis is our own):

- **Surveys or case studies on Semantic Web technology in enterprise systems**
- Comparative studies on the evolution of Semantic Web adoption
- **Semantic systems and architectures of methodologies for industrial challenges**
- **Semantic Web based implementations and design patterns for enterprise systems**

- **Enterprise platforms using Semantic Web technology as part of the workflow**
- Architectural overviews for Semantic Web systems
- Design patterns for semantic technology architectures and algorithms
- System development methods as applied to semantic technologies
- Semantic toolkits for enterprise applications
- **Surveys on identified best practices based on Semantic Web technology**
- Linked Data integration and change management

7.7 Workshop on Linked Data Quality

The Workshop on Linked Data Quality, collocated with SEMANTICS 2014, mentioned the following topics of interest in their call for papers (emphasis is our own):

- Approaches targeting linked data in the areas of:
  - **quality assessment**
  - inconsistency detection
  - **cleansing, error correction, refinement**
  - versioning
- **Reputation and trustworthiness of web resources**
- Quality of ontologies
- Quality modelling vocabularies
- **Frameworks for testing and evaluation**
- **Data validators**
- Best practices for linked data management
- User experience
- Empirical studies

7.8 1st Workshop on Linked Data for Knowledge Discovery (LD4KD)

The 1st Workshop on Linked Data for Knowledge Discovery (LD4KD), collocated with ECML/PKDD, included the following topics of interest in the call for papers:

- Linked data for data pre-processing: cleaning, sorting, filtering or enrichment
- Linked data applied to machine learning
- Linked data for pattern extraction and behaviour detection
- Linked data for pattern interpretation, visualization or optimization
- Reasoning with patterns and linked data
- Reasoning on and extracting knowledge from linked data
- Linked data mining
- Link prediction or links discovery using knowledge discovery and data mining
- Graph mining in linked data
- Interacting with linked data for knowledge discovery

7.9 Linked Open Data 2014: Improving SME Competitiveness and Generating New Value

The Workshop on Linked Open Data: Improving SME Competitiveness and Generating New Value, collocated with SEMANTICS 2014, included the following topics in their call for papers (emphasis is our own):

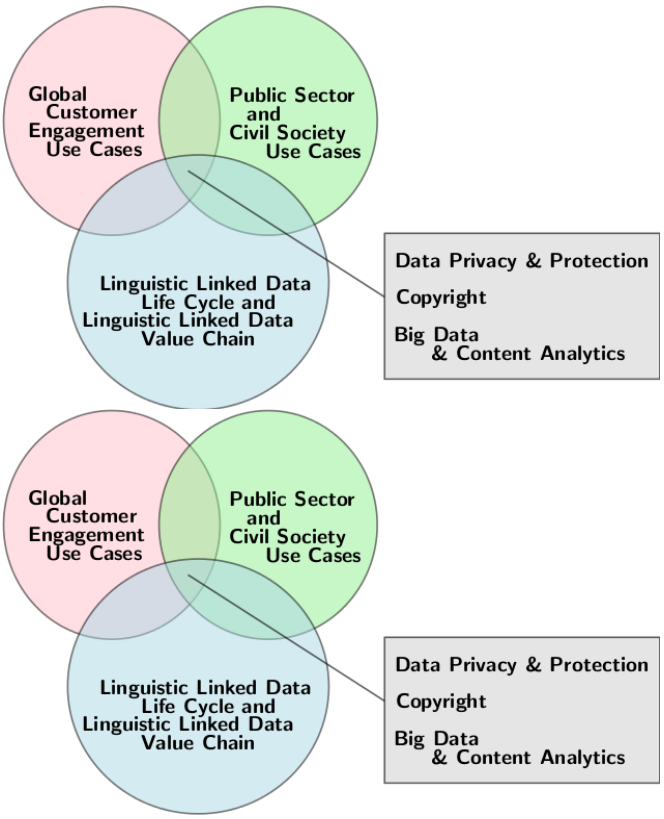
- **Linked data for SMEs**
- **Managing the data Life cycle in SME environments**
- Analytics for improving business knowledge using linked data
- Transforming data to open and linked formats
- Business collaboration through data sharing and alignment

7.10 Summary

Research in linked data is increasingly **focusing on issues related to the consumption and application of linked data, also in commercial and SME contexts**. Key aspects related to the consumption and exploitation of linked data include i) **description and discovery of linked data**, ii) **validation and quality assurance** and iii) **provenance, privacy and IP management**. The research community is thus increasingly working on the above mentioned issues and will eventually contribute to a linked data **ecosystem where data trust, provenance and licensing information is taken into account and appropriately represented**, and infrastructure is available to ensure for the **low-cost publishing, discovery, validation and reuse of linked datasets**. Further, current efforts consider how SMEs can exploit linked data technology as part of their workflows and in enterprise applications.

8 Roadmap

We structure the roadmap into three key application areas: i) **Global Customer Engagement Use Cases**; ii) **Public Sector and Civil Society Use Cases**; iii) **Linguistic Linked Data Life Cycle** and **Linguistic Linked Data Value Chain.**
We also discuss aspects that are at the intersection of these three application areas, including licensing, copyright and the ability to process big datasets and perform multilingual content analytics, which are relevant to all application areas.
The following figure provides a graphical representation of the three application areas:



The goal of the LIDER project is to identify actions and developments that need to happen in order for linguistic linked data technologies to impact the three fields identified above. We frame our predictions by indicating the following dimensions:

- **Horizon**: Distinguishing between 1-2 years, 3-5 years, and 5-10 years from the publication of this report, with the horizon of 1-2 years corresponding to initiatives that have already been started (e.g. European research projects funded under FP7 or started in 2014 under Horizon 2020), 3-5 years corresponding to initiatives planned but not started yet, e.g. as foreseen in the Horizon 2020 work program. The horizon 5-10 years corresponds to activities that are not yet foreseen in any plan or work program but are expected to emerge.
- **Main actors**: Identifying the main actors who are involved in the initiative and who will push it. We distinguish between the following actors: academia, industry, SMEs as well as partnerships between them.
- **Means**: Instruments by which the initiative can be realized, e.g. collaborative research projects, cooperation between academia and industry, standardization, industrial pilots, mainstream adoption, and academic proof-of-concept.

8.1 Global Customer Engagement Use Cases

8.1.1 Media Publishing and Content Management

As identified by Forrester (Yakkundi et al. 2013), one of the key challenges in media publishing and content management for the years to come lies in the convergence of technologies to deliver a homogeneous multichannel experience to customers, or as stated explicitly in the report by Forrester: *"The dynamic nature of digital experiences requires technology that enables business users to manage, measure, and optimize what happens on web, mobile, and social channels".* The key objective to achieve here is to **combine different technologies to provide a contextualized digital experience to users, bringing together content, product and customer relationship management in one ecosystem**. Portals are expected to continue playing an important role in delivering a personalized environment that supports digital customer experiences. The important issue is to **realize a certain degree of agility at the content level to be able to quickly integrate new (external) data resources** to react to new needs and interests of customers.

The key impact avenues for linked data based content analytics in media publishing and content management are thus the following ones:

- Relying on linked data technologies to **create unified information spaces of linked datasets bringing together datasets that have been isolated so far** (e.g. product content, product data, customer data, social data etc.) to **contribute to a unified experience** (so called "*vertical clouds"*)
- Exploiting **terminologies and ontologies** available as linguistic linked data **to achieve terminological consistency across channels and languages**
- **Agile import of datasets** into portals to support changing customer needs and interests
- Support for **localization of content across channels** by exploiting multilingual linguistic linked data resources

We expect that in the future linked data technologies for content publishing will mature to make multimodal and multilingual **repurposing of content and storytelling feasible and practical, and to lower the cost for doing so**. To this end, it will be crucial to **support access to knowledge, and not only data, by non-experts**, e.g. content creators and consumers, developing interfaces that abstract from technical aspects, data models and query languages. This access should in particular be across media and across natural languages. This will ultimately lead to **visual story generation from multiple sources**

including text, video and other modalities as well as new methods for re-purposing and composing heterogeneous content for different challenges, natural languages and audiences.

We need further **best practices for linked data based content publishing as well as experience reports on the adoption of such best practices in verticals** such as energy efficiency management, smart cities, and healthcare, together with **best practices and business models for generating value out of such resources**. In particular, we advance that enterprises will realize in the coming years the **potential of linked data to connect different media beyond separate annotations (cross-media links) and the potential of linked data to share such data across companies.**

The integration of content comprises the **creation of a seamless network of data and knowledge that spans multiple modalities as well as open and closed datasets in a way that is respectful to intellectual property (IP) and corresponding licenses**. Assuming that in the future machines will be mainly responsible for discovering and mashing up content from heterogeneous sources in different formats, modalities and languages, methodologies that ensure that copyright, IP, confidence and provenance of data are taken into account need to be developed. Technically, this will require linked data aware licensing servers which will be responsible for delivering information taking into account access rights, supporting machine-mediated access, extraction, aggregation, composition and repurposing of data.

A further challenge will be to **deploy linked data technology to capture social interactions at large scale**, annotating multiple media streams with respect to who says what to whom, and what reactions are triggered by which content. This includes **capturing emotions across modalities**, bringing different modalities and languages together to analyze emotions and sentiment effectively.

In the long term we foresee that **linked data will converge to create a seamless ecosystem in which structured and unstructured information from different modalities and languages will be integrated and linked**, thus being exploitable by a new generation of methods and services that will support storytelling and question answering over all these sources. This assumes that the technology stack for analyzing content using natural language processing techniques at large scale has matured enough.

Supporting exploration of data by citizens and the larger public including analysts, journalists and the like is another challenge. This will effectively contribute to dealing with data and information overload.

Using terminologies and ontologies available as linguistic linked data to achieve terminological consistency across channels would certainly contribute to exploiting data in commercial contexts.

Roadmap for Media Publishing and Content Management

Horizon	Prediction	Actors	Means
1-2 Years	Increased linked data publishing in verticals and for different use cases, first ROI models emerge.	Industry-Academia Partnerships	Productive Systems and Integrated Projects
1-2 Years	Increased multilingual terminologies and ontologies in verticals and for different use cases and channels, first ROI models emerge.	Industry-Academia Partnerships	Productive Systems and Integrated Projects
1-2 Years	Effective solutions and interfaces for access to data for non-expert content creators are in place.	Academic and Industrial Cooperation	Pilots
1-2 Years	Increased awareness about potential of linked data to connect different media beyond monomodal annotations (cross-media links).	Start-ups and Academia	Hand-on Workshops, Tutorials, Webinars
3-5 Years	Robust and accurate techniques for linking ontologies across languages are available and successfully applied.	Industry	Productive Systems
3-5 Years	An increasing number of media publishing companies publish their data about programs,	Industry	Productive Systems

	background information as part of the Web, following the early adopters such as the BBC.		
3-5 Years	Multilingual and cross-media access and exploration of data by citizens becomes possible.	Industry-Academia Partnerships	Research Projects and Pilots
3-5 Years	Techniques for bringing together data from multiple modalities for holistic sentiment analysis mature and become productive.	Industry	Productive Systems
3-5 Years	Standards for annotation and exchange of multimodal datasets emerge.	Industry-Academia Partnerships	Research Projects and Pilots
3-5 Years	Methods for multimodal storytelling by repurposing, summarizing and composing existing and heterogeneous multimodal content are available.	Industry-Academia Partnerships	Research Projects and Pilots
3-5 Years	First solutions to combine open and public datasets with appropriate handling of IP and licenses emerge as well as first solutions to discover and assess trust and quality of linked data.	Industry-Academia Partnerships	Research Projects and Pilots
3-5 Years	Techniques for large-scale capturing of social interactions and their semantics across modalities emerge	Industry-Academia Partnerships	Research Projects and Pilots
5-10 Years	Non-public data will be available as linked data on the Web; linked data aware licensing servers take into account and reason about access rights in delivering content.	Industry-Academia Partnerships	Research Projects and Pilots
5-10 Years	Linked data technology supports the seamless integration of structured and unstructured data available on the Web as well as querying across languages of this integrated data.	Industry-Academia Partnerships	Research Projects and Pilots
5-10 Years	Linked data based multimodal and multilingual storytelling matures and is adopted.	Industry	Productive Systems

8.1.2 Marketing and Customer Relationship Management

Extrapolating from the importance of social aspects described in Section 4 and the needs expressed by the survey carried out by Alta Plana (see Section 4), our own surveys (LIDER project deliverable D1.1.1), as well as the findings from the 4th LIDER roadmapping workshop), we can assume that the **analysis of the voice of the customer** will play a major role in guiding marketing and advertising activities in the future. What will be needed in the future are **robust techniques to extract and interpret the voice of the customer** with i) **a high level of accuracy**, ii) **across natural languages and modalities**, and iii) **analyzing sentiment at deeper levels beyond mere polarity** in order to also recognize the intent of a user as a basis to generate actionable knowledge. The insights generated by methods to analyze the voice of the customer need to be converted into appropriate metrics that can be integrated into standard BI solutions to be correlated with other measures and in order to measure the impact and ROI of a certain marketing campaign. The holy grail of the advertising industry is to aggregate data from potential and actual customers ubiquitously as a basis to create deep personal profiles in order to provide personalized and contextualized recommendations that permeate the whole life of a user. An important challenge herein is to be able to process large amounts of Big Data in real time (see Section 8.4.3). With respect to creating deep personalized profiles of users, **domain-specific background knowledge available as linked data can be exploited to create rich semantic profiles that support contextualized, personalized and situated recommendation and interaction**. **Linked data technologies can also play a key role in linking profiles of users across channels and sites**. This needs to take into account the right of people for privacy (see Section 8.4.1).

In particular, linked data can contribute to this challenge as follows:

- **Sentiment lexica in different languages**, including polarity information and link to intentions are available as part of the linguistic linked open data (LLOD) cloud, so that

these lexica are easily integrateable into standard sentiment analysis tools and workflows
- **Ontologies modeling intentions for a number of micro-domains** become part of the LLOD
- **Datasets annotated with sentiment, subjectivity, polarity and potentially irony** for many languages become part of the LLOD
- **Robust methods for linking and identifying users across channels** and sites as a basis to create aggregated user profiles
- **Robust and accurate methods for detecting sentiment, subjectivity and polarity and even irony for the major European languages** as LLOD-aware services
- Providing **ontologies/taxonomies/terminologies** that can be used to represent **semantic profiles of users**
- Providing **ontologies for modeling situations and contextual parameters to provide situated and contextualized recommendations**
- Providing **domain-specific terminologies and ontologies lexicalized in multiple languages and across modalities to provide consistency across channels, languages and modalities**

In general, we can expect disruptive changes to the current paradigm for marketing and customer relationship management. Marketing and advertising activities can be assumed to become totally transparent to the user, moving from an active push over advertising strategies that are embedded in social conversations and communications to the recognition and fulfillment of intentions and needs in real time. Explicit marketing and advertising activities will lose importance as customer targeting and product placement becomes a commodity service that is perceived as a real added value by customers. This has several implications:

- **Machine-to-machine communication:** As advertising and marketing moves from a push to a commodity service that recognizes and fulfills customer needs in real time, we can expect that both businesses and consumers will move from being real physical entities to being digital agents that interact and negotiate directly. This will radically change current business models.
- **Rich semantic user profiles:** Real-time recognition of needs and intentions requires rich linked information including semantic information about objects, individuals, groups, intentions, contexts, cultures, etc. This will require standardized ways for representing and linking such information.

As more and more private data is linked, profiles become more and more important, and **online reputation becomes increasingly relevant for online transactions**, new business models will emerge, such as offering to manage online profiles and reputation.

We also foresee that advertising and product placement will become a commodity service that fullfills needs in real time, thus becoming completely transparent, personalized and contextualized. However, this will lead to a situation in which **users experience a lack of control and most likely to a pushback in which users demand for more control of their personal data**. Consequently, this will require new solutions and paradigms to empower users in revoking their personal data and solutions for unlinking datasets, thus giving back control to users over their personal data (see Section 8.4.1)

Roadmap for Marketing and Customer Relationship Management

Horizon	Prediction	Actors	Means
1-2 Years	Paradigm shift from physical interaction with real customers to machine-to-machine interaction.	Industry	Paradigm Shift
1-2 Years	Paradigm shift away from CRM as a push activity to a (moderated) conversation.	Academic and Industrial Cooperation	Paradigm Shift
1-2 Years	Need for standardized vocabularies for describing user profiles, product information and their relations.	Academia-Industry Partnerships	Standardization activities
3-5 Years	Robust and accurate techniques for linking terminologies and ontologies across languages are available and successfully	Industry	Productive Systems

	applied.		
3-5 Years	Advertising industry exploits rich semantic interlinked personal and product profiles to provide personalized and contextualized recommendations.	Industry	Productive Systems
3-5 Years	First solutions for centralized and trusted management and storage of personal data emerge that consider provenance and licensing terms and conditions and ensure compliance with these.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Paradigm shift moving away from explicit advertising to transparent advertising, recognizing intentions and needs in real time to fulfill them, becoming a commodity beyond mere recommendations.	Industry-Academia Partnerships	Research Projects and Pilots
3-5 Years	Services for personal data tracking and revocation start to become available.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	New business models emerge, e.g. for online reputation management and optimization of individuals, companies etc.	Start-ups, Academia	Pilots
5-10 Years	Personalization in advertising is increasingly perceived as a threat by end users and as interfering with their free choice, demand for more control over data arises; advanced solutions allowing users to manage their personal data arise.	Industry-Academia Partnerships	Research Projects and Pilots
5-10 Years	Methods and best practices for retracting and unlinking personal data emerge.	Industry-Academia Partnerships	Research Projects and Pilots

8.2 Public Sector and Civil Society Use Cases

8.2.1 Supporting the Creation of a Single Digital Market and the Connecting Europe Facility (CEF)

One of the main stated goals of the Connecting Europe Facility (CEF) is to **overcome service fragmentation and lack of interoperability due to national borders**, with the objective to develop **pan-European services that interoperate across borders and defragment the market**, in particular in the areas of eGovernment, eProcurement and eHealth.

For this, **datasets exchanged across borders need to be harmonized both at a syntactic and at a semantic level**. While interoperability at the syntactic level is being addressed already to some extent, establishing interoperability at the semantic level involves a long-term effort involving the alignment of concepts used in different countries and jurisdictions.

With its strong tradition working on ontology alignment and linking, the semantic technologies community has the potential to contribute to:

- The **development of shared ontologies of key administrative and legal concepts across Europe**
- **Linking of vocabularies and ontologies existing in different countries and jurisdiction to foster interoperability**
- Development of **declarative specifications of workflows and processes**, so that tools can reason about them and compose them to achieve some task
- **Collaborative ontology creation across languages and countries**
- **Exploitation of terminologies and ontologies to ensure consistency of communication** in public administration

Roadmap for Supporting the Creation of a Single Digital Market and the Connecting Europe Facility (CEF)

Horizon	Prediction	Actors	Means
1-2 Years	Shared ontologies and terminologies for key administrative, financial and legal sub-domains emerge as part of the linguistic linked open data cloud; these ontologies are lexicalized in multiple languages following standard vocabularies and best practices such as the lexicon model for ontologies.	Academia-Industry Partnerships	Research Projects and Pilots
1-2 Years	Ontologies are increasingly linked across national contexts and published on the linked open data cloud.	Academic and Industrial Partnerships	Research Projects and Pilots
1-2 Years	Most important domains in which terminological standardization across languages is needed to realize the vision of a single digital market are identified.	Academic and Industrial Partnerships	Research Projects and Pilots
1-2 Years	A taxonomy of types of links that matches needs of the single digital market is developed.	Academia-Industry Partnerships	Standardization activities
1-2 Years	Strive to reduce the amount of unstructured data exchanged by exploiting cross-lingual ontologies and terminologies that diminish the need to exchange unstructured messages; identify those fields where language-independent communication is applicable and feasible.	Academia-Industry Partnerships	Standardization activities
3-5 Years	First standards and best practices including ontologies to describe services and products emerge that can be exploited in machine-to-machine negotiation and matchmaking (e.g. matching job offers to profiles).	Industry	Standardization activities and Pilots
3-5 Years	Robust and accurate techniques for linking ontologies across languages are available and successfully applied.	Industry	Productive Systems
3-5 Years	The need to formally monitor compliance of actors across countries with European and other regulatory frameworks becomes obvious, and first solutions for expressing policies and regulations become available (e.g. using advanced logics such as deontic logics).	Academia	Research Projects
5-10 Years	A rich ecosystem of reference and localized ontologies describing key domains in which data exchange across language and national boundaries is crucial has emerged together with an ecosystem in which validated and trusted mappings at different levels of trust and provenance are available.	Industry-Academia Partnerships	Research Projects and Pilots
5-10 Years	Robust methodologies for collaborative development of shared ontologies across cultural contexts emerge and are successfully applied.	Academia-Industry Partnerships	Research Projects and Pilots
5-10 Years	Solutions for monitoring compliance of data and services to regulations are available and deployed at a large scale.	Industry	Productive Systems

8.2.2 Localization and Translation

Linked data technologies have the potential to impact the current localization and translation market and processes by providing more flexible ways of publishing and exploiting multilingual datasets including parallel corpora, terminologies and translation memories, but also other multilingual datasets of common interest (e.g. DBpedia and BabelNet). **Best practices and standards for publishing parallel texts as part of the**

linguistic linked data cloud and their exploitation in standard localization and translation workflows are needed.

Linked data has the potential to be exploited by translators but also content creators. **New paradigms in which content creation and translation are intertwined in the sense that machine translation can be exploited in bootstrapping content creation and vice versa will become feasible in the near future.** Standards and best practices for licensed linked data need to be developed so that the localization industry can trustfully work with linked data. In the long-term, we foresee that **linked data will be an enabler of high-quality personalized translation**.

Translation of terminologies and speech-to-speech translation will be two important application fields of machine translation. **Translation of terminologies** is crucial to realize the idea of a single digital market and in order to ensure **terminological consistency across players and stakeholders and national sites. Speech-to-speech translation is going to play a key role in content creation and consumption**, allowing to translate educational offerings, social applications, TV programs and so on in real time.

In the longer term, a crucial issue will be to **extend coverage to non-European languages** and extend the experiences to the languages and countries in which these languages are spoken.

Some of the above mentioned aspects have been identified as relevant as part of the 3rd LIDER Roadmapping Workshop collocated with Localization World 2014 in Dublin (also see LIDER Deliverable D4.6).

Roadmap for the Role of Linked Data in Translation and Localization

Horizon	Prediction	Actors	Means
1-2 Years	Linked data is increasingly exploited as background knowledge and context by translators and content creators.	Industry	Productive Systems
1-2 Years	Best practices and standards for publication of parallel texts, terminologies and translation memories as linked data emerge.	Academic and Industrial Partnerships	Standardization
1-2 Years	Best practices and standards for publication of speech-to-speech translation data emerge and are increasingly applied.	Academic and Industrial Partnerships	Standardization
3-5 Years	A synergetic feedback loop between translation and content creation is developed in many working systems, thus bootstrapping multilingual content creation by MT services and feeding back corrections to the MT services to improve in the longer term.	Industry & Non-profit organizations	Productive Systems
3-5 Years	Localization industry is increasingly aware of linked data; licensed linked data solutions make it possible for localization industry to share and exploit linked data content.	Academia-Industry Partnerships	Standardization activities
3-5 Years	First solutions for live updates of resources by online communities emerge that take into account IP, licensing and provenance aspects adequately.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Large-scale terminology translation and alignment contributes to establish a single digital market across EU member languages.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Robust techniques for speech-to-speech translation emerge and are deployed in commercial settings and applications.	Industry	Productive Systems
5-10 Years	Terminologies for relevant domains are standardized, supporting automatic consistency checks across legislations become possible.	Academia	Research Projects
5-10 Years	Fully personalized and contextualized translation supported by linked data is deployed in commercial systems and existing localization workflows.	Industry	Productive Systems

8.2.3 Open Data Commons, Data Quality and Data Lifecycle

Open data can generate great value by allowing third parties to improve services (Pentland 2014). This vision is certainly compatible with the ideas behind the linked open data project. Alex Pentland has in particular stated that *"A key insight is that our data are worth more when shared because they can inform improvements in systems such as public health, transportation, and government. Using a "digital data commons" can potentially give us unprecedented ability to measure how our policies are performing so we can know when to act quickly and effectively to address a situation."* Such a creative commons can be used as a basis to optimize important aspects of our life, including transportation, traffic, energy networks and it can also build the basis for the vision of smart cities (Komninos 2009) to improve local and regional governance.

A crucial issue in building and using data commons is ensuring **high-quality and up-to-dateness of the data available as part of the data commons**. This requires methods for ensuring quality of datasets over their whole lifecycle, monitor usage etc. but also to **ensure appropriate access control if data access has to be restricted**. In addition, it will be important that **all data parts of the data commons are enriched by corresponding licensing information that states the terms and conditions under which the resource can be used, and for which purposes**. Such licensing information needs to become an integral part of the data and be in machine-readable format in order to allow for automated discovery, reasoning and filtering. In building a data commons, it is important that datasets are linked across languages to support cross-country and cross-lingual comparisons.

As linked data and open data becomes more central to the development of applications, there will be a focus on ensuring that the quality of this data is maintained, and that the tedious process of **data wrangling is reduced**. This will take the form of services that provide **validation and certification** of datasets on both the syntactic correctness as well as the semantic correctness of the data. Such certification will allow data consumers to trust data producers, while the data producers in turn get direct feedback on the quality of their data. Furthermore, services will increasingly be created that **augment the data automatically** by means of automatically adding new information or by aggregating and linking data from multiple sources.

Roadmap for Supporting the Open Data Commons

Horizon	Prediction	Actors	Means
1-2 Years	Linked Data Network is considered a viable option for the realization of data commons.	Academia-Industry Parternships	Research Projects & Pilots
1-2 Years	Services for automatic validation and certification of datasets become available.	Academia-Industry Parternships	Research Projects & Pilots
3-5 Years	Most linked data endpoints implement an access control layer.	Academia-Industry Partnerships	Research Projects & Pilots
3-5 Years	Automatic validation services cover all data published in meta-data repositories automatically.	Academia-Industry Partnerships	Research Projects & Pilots
3-5 Years	Automatic data alignment and transformation becomes available.	Academia-Industry Partnerships	Productive Systems
3-5 Years	Services for monitoring quality, evolution and uptime of datasets as part of the data commons are in place, mature and widely used.	Academic and Industrial Partnerships	Research Projects & Pilots
5-10 Years	Data consumption and linking is mostly performed without manual intervention.	Industry	Productive Systems

8.3 Linguistic Linked Data Life Cycle and Linguistic Linked Data Value Chain

8.3.1 Linguistic Resource Development and Sharing

Many of the use cases and needs mentioned in the sections above require the availability of linguistic data, corpora and lexical resources in multiple languages. Such resources are generally scarce, as identified by MetaNet.

Towards increasing the coverage and quality of linguistic resources, the following aspects need to be considered:

- **Licensing information and provenance**: Licensing and provenance information need to be attached to the data, ideally in machine-readable form, defining the conditions under which data can be used, also in commercial settings in which revenue is obtained from the use of the resource.
- **Resource market**: A business-to-business and research-to-business market for high-quality data trading needs to emerge; one possibility is to market data per API and establish pay-per-use models.
- **Quality and availability**: Resources need to have high availability as well as high quality. Frameworks for monitoring availability and quality of resources need to be established.
- **Stimulation of resource development**: The creation of resources can be stimulated by initiating partnerships between academic and industrial consortia that jointly work on a dataset that all can exploit for their purposes at zero cost under the condition that the resource is licensable by third parties under fair conditions.
- **Sharing and discovery**: The easy sharing and discovery of resources needs to be ensured by appropriately extending the functionality of current metadata repositories for linguistic data.

The above requirements were to a great extent identified during the 1st LIDER roadmapping workshop collocated with the European Data Forum (EDF) in Athens (also see LIDER Deliverable D4.5).

Roadmap for Language Resource Development and Sharing

Horizon	Prediction	Actors	Means
1-2 Years	Vocabularies for providing licensing information in machine-readable form are standardized.	Academia-Industry Partnerships	Standardization
1-2 Years	Vocabularies for describing linguistic datasets become standardized.	Academia-Industry Partnerships	Standardization
3-5 Years	Metadata repositories implement improved functionalities for discovery of relevant resources by means of current Web standards (e.g. SPARQL); these repositories adopt the standardized vocabularies developed by relevant communities in addition to their own vocabularies based on internal data models to allow for interoperability.	Academic and Industrial Partnerships	Research Projects and Pilots
3-5 Years	Metadata repositories ensure that the link to the actual data is available, following a number of best practices.	Academic and Industrial Partnerships	Research Projects and Pilots
3-5 Years	Agent-to-agent negotiation for data exploitation is realized in pilots.	Academia-Industry Partnerships	Pilots
3-5 Years	Aggregator services that collect, aggregate and index metadata about linguistic resources emerge providing added value as brokers and are increasingly used to discover linguistic resources.	Academic and Industrial Partnerships	Research Projects and Pilots
5-10 Years	A cross-border market for linguistic resources is established and in operation.	Industry, Academia and Non-profit organizations	Productive Systems
5-10 Years	An ecosystem of services that validate and benchmark data are available and widely used.	Industry, Academia and Non-profit organizations	Productive Systems

8.3.2 Linguistic Linked Data Value Chain

D3.2.1 Roadmap for the use of Linguistic Linked Data for content analytics – Phase I

A crucial aspect in the future will be the **establishment of a value chain and appropriate ecosystem and infrastructure for the creation, marketing, exchange, consumption and modification of linguistic data.** The linguistic linked data cloud can provide the basis for such an ecosystem, but the following aspects need to be taken into account:

- **Business models**: Business models for all the actors along the chain need to be developed, i.e. for resource creators, resource traders or brokers including discovery platforms, as well as those actors that enrich, manually validate and improve datasets; non-monetary rewards and transactions in terms of community recognition or increase of reputation need to be explored.
- **Trust and rating**: An ecosystem for assessing trust and rating stakeholders based, e.g., on reputation is needed.
- **Data quality and benchmarking**: Methods for independent measurement and benchmarking of data quality need to be developed. Quality models for comparing different resources are needed. It is important to stress that data quality is nevertheless difficult to measure inherently and to a large extent is determined by the value that the data generates in terms of ROI for certain applications and is thus something that the market should agree upon. Data quality is in many cases also determined by factors external to the data itself, i.e. by documentation, recommendations of data by others etc.
- **Standardization and plurality**: Standardized formats for resources, services and APIs need to be agreed upon as a community effort and based on open standards, vocabularies and best practices. At the same time, plurality and backwards compatibility needs to be supported, having converters from legacy formats into standard formats available as part of the value chain.
- **Semantic Web extensions**: Extensions to SPARQL and other Semantic Web technology will be needed, as SPARQL does not perfectly match all use cases, especially for querying parallel text and speech data.
- **Data curation and improvement**: Community contributions to improve a certain resource are in line with the distributed nature of linked data, but infrastructure, tools and workflows to support this need to be developed.
- **Call for tenders**: Call for tender models in which potential customers can call for the development of a certain resource will likely play an important role in the future.
- **Integration:** Plug and play integration of external datasets with internal datasets should be supported, mashing up and combining datasets should be supported by agreed-upon common formats and appropriate tooling.
- **Architecture for LLOD-aware services:** Service architectures that scale and rely on distribution and stream processing and exploit open Web protocols and Semantic Web standards (e.g. JSON-LD) to provide content analytics services and exploit the LLOD as background knowledge are needed.

Roadmap for Linguistic Linked Data Value Chain

Horizon	Prediction	Actors	Means
1-2 Years	New business models for linguistic resource creation also based on non-monetary transactions and rewards emerge.	Academia-Industry Partnerships	Research Projects and Pilots
1-2 Years	Proposal for an architecture for LLOD-aware services building on open Web and Semantic Web standards emerge.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Business models for data creation, brokering and improvement are established.	Academia-Industry Partnerships	Productive Systems
3-5 Years	Models for the creation of linguistic resources following the call for tenders paradigm are explored in pilots.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	First quality models for linguistic linked data emerge.	Academia-Industry Partnerships	Pilots
3-5 Years	Principles and best practices for measuring and representing trust and quality of datasets and services by independent parties that provide independent certification services for linguistic resources emerge.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Best practices and standards for publishing the most important types of linguistic resources are available and widely used; converters to migrate legacy data into the LLOD cloud are available.	Academia-Industry Partnerships	Research Projects and Pilots

5-10 Years	Models based on call for tenders for the development of linguistic resources are in place and widely used.	Industry, Academia and Non-profit organizations	Productive Systems
5-10 Years	LLOD-aware services following architectural best practices and open Semantic Web vocabularies and Web protocols that can be easily composed and integrated into worfklows are widely deployed.	Industry, Academia and Non-profit organizations	Productive Systems

8.4 Orthogonal Topics

8.4.1 Data Privacy and Data Protection

As more and more (personal) data is collected from users, the crucial issue how to ensure responsible and trustful handling of this data emerges. In this respect, the report *Rethinking Personal Data: A New Lens for Strengthening Trust* published by the World Economic Forum mentions that *"the growth of data, the sophistication of ubiquitous computing and the borderless flow of data are all outstripping the ability to effectively govern on a global basis"*, with the result that *"industry, government and civil society are all uncertain on how to create a personal data ecosystem that is adaptive, reliable, trustworthy and fair"* (Kearney 2014).

Especially critical here is passively generated data, e.g. data generated by sensors, devices or wearables, as users are typically unaware of the data generated nor have they typically provided consent for the use of the data.

In fact, the World Economic Forum has started a global dialogue on the topic of how to ensure responsible use of personal data, identifying the following three key issues:

- **Delivering meaningful transparency:** Giving individuals higher transparency with respect to how data is used, simplifying the way in which data practices are communicated to individuals.
- **Strengthen accountability:** Creating an ecosystem and incentive infrastructures to ensure principled and enforceable data use. This implies that there needs to be verifiable evidence by stakeholder organizations that relevant measures are being taken to ensure compliance with data usage best practices.
- **Empower individuals:** Individuals should be empowered to be able to decide about how their data is used but also to be able to use their data for their own purposes. Moreover, they should be able to understand and manage the impact of data usage.

It has been argued that a first step in developing good practices for responsible and trustful data usage is a taxonomy of types of personal data (Kearney 2014, Chapter Near-Term Priorities for Strengthening Trust) that distinguishes at least the following kinds of data:

- **Individually provided data:** data voluntarily provided by individuals through forms, surveys, social media applications etc.
- **Observed data:** data generated by some mobile or wearable device or any other sensor
- **Inferred data:** data generated as part of some data mining or machine learning algorithm on the basis of either individually provided or observed data

While the user is aware and actively participates in the provision of individually provided data, this is not the case for observed or inferred data where there is typically no awareness from the side of an individual on which data is being collected and for which purpose.

A reference model for personal data management has been outlined in (Kearney 2014, Chapter Long-Term Issues and Insights); it distinguishes the following three layers:

- **Infrastructure:** The infrastructure layer comprises the technology, services and applications required to assure the availability, confidentiality, security and integrity of the data, both while in transit and at rest.
- **Data management:** The data management layer focuses on the transfer and exploitation of personal data as specified in corresponding permissions and policies. Metadata is crucial to enrich the data by a layer that allows to express permissions and provenance as a basis to ensure compliance with agreed-upon policies and terms and conditions. The metadata regarding licensing and provenance needs to remain attached to the data during the whole data lifecycle.
- **User interaction:** The user interaction layer facilitates a transparent interaction of individuals with service providers regarding the terms and conditions of their personal data.

In addition, Alex Pentland (Pentland 2014) mentions a set of five key policy recommendations for large organizations with respect to data management:

- **Distribution:** Large data systems should store data in a distributed manner, separated by type (e.g. financial vs. health) and real-world categories (e.g. individual vs. corporate).
- **Provenance and views:** Data sharing should always maintain provenance and permissions associated with data, and should support automatic, tamper-proof auditing. Best practice here would be to share answers only to questions about the data, e.g. by relying on views rather than sharing the data themselves, whenever possible.
- **Secure external data sharing:** External data sharing should take place only between data systems that have similar local control, permissions, provenance, and auditing, and should include the use of standardized legal agreements such as those employed in trust networks.
- **Best practices for data flows:** Best practice for data flows to and from individual citizens and businesses is to require them to have secure personal data stores and be enrolled in a trust network data sharing agreement.
- **Secure identification protocols:** All entities should employ secure identity credentials at all times.

In a data economy in which data is the new oil and many processes and services are optimized in a data-driven fashion and thus large amounts of data, also pertaining to individuals, is accumulated, the issue of how to ensure trust and preserve privacy is of utmost importance. Developing models that allow to create value out of data while at the same time respecting privacy rights is thus an important topic on the political agenda.

The center of the European agenda of actions to improve the data protection and the privacy of EU citizens is the General Data Protection Regulation (GDPR) expected to be issued in 2014. This regulation extends the scope of the EU data protection law to all foreign companies processing data of EU residents, harmonizes the data protection regulations throughout the EU (supported by the European Data Protection Board) and proposes a single, centralized Data Protection Agency to be responsible for taking legally binding decisions against a company (GDPR, 2012). In the new regulation, the notice requirements will be strengthened, and additional information will have to be provided regarding the retention time for personal data, the contact information of both the data controller and the data protection officer. Data controllers will have to prove that they have the person's consent and will have to inform faster if a data breach has occurred, informing also the affected person. The right to erasure (Art. 27) will improve the user’s privacy (see the case affecting Google, ECLI:EU:C:2014:317) but will also add a burden on the data management procedures.

There are several technical measures that have been identified as clearly contributing to the creation of a trustful data ecosystem that respects the privacy rights of individuals and to which linked data technologies can clearly offer a contribution:

- **Distribution**: Partitioning data and physically distributing it across sites is a measure to make aggregation of personal data more difficult and thus ensure that no single individual has access to all datasets about a person in one place (Pentland 2014). Linked data can strongly contribute to this as it inherently relies on data distribution across a network of physically distributed but logically linked datasets. **Linked data could thus provide the basis for data distribution** as envisioned by Alex Pentland (Pentland 2014).
- **Access control**: Appropriate access control and authentication services are needed to ensure that data access is limited to those users that are authorized. The linked data technology stack thus needs to be extended by an access control and authentication layer.
- **Views**: Alex Pentland argues that most data access should be in the form of **views** rather than through direct access to the data. This is also compatible with linked data technologies where SPARQL can be seen as a language to define views and expose the results of this view over the Web. Combined with appropriate mechanisms for access control, linked data could thus provide the basis for the view-based access to data that Pentland advocates.
- **Machine-readable provenance and licenses:** It is crucial to **ensure compliance with licenses and that license and provenance information are attached to the data and remain so during the whole data lifecycle** in order to prevent misuse of the data. It is crucial to **make the terms and conditions specify how the data can be used transparently in all steps of the data lifecycle**. Machine-readable licenses that can be attached to the data (e.g. in RDF) are thus needed and could be made an integral part of the data.

8.4.2 Copyright

Much of the linguistic data is indeed subject to the Intellectual Property laws. Regardless of whether the language resource is expressed as linked data or not, most of the published resources have a protection that has to be respected by data consumers. This protection comes from the EU Copyright Directive (Directive 2001/29/EC) if the resource is a work, and from the EU Database Directive (Directive 1996/9/EC) if the resource is simply regarded as data.

Even if a large percent of the published material is published as open data (namely, with light requirements imposed by the rightsholder for the resource to be used, derived or redistributed), a non-negligible portion of the resources is published with strong restrictions on its use.

The landscape is likely to be modified in the forthcoming years. A public consultation on the review of the EU copyright rules was held in 2013 and 2014 and mid-term changes in EU copyright law may be on the horizon. These changes will probably integrate a better handling of digital resources and their licenses published online, more uniform rules throughout Europe and a more transparent management of the copyright collecting societies. This may imply forcing a swifter management of the rights and an increased importance of automated rights management systems.

Roadmap for Privacy and Trust

Horizon	Prediction	Actors	Means
1-2 Years	Regulation on the processing of personal data (GDPR) are in force.	Legal Bodies	New Regulation
1-2 Years	Proposals for access control over linked data are consolidated and standardized.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Most linked data endpoints implement access control layer.	Academia-Industry Partnerships	Productive Systems
3-5 Years	A new copyright directive empowers automated right management processing.	Legal Bodies	A New Directive
5-10 Years	Standards for provision of licensing and provenance information in RDF are available.	Academia-Industry Partnerships	Standardization
5-10 Years	Inclusion of licensing and provenance information in	Industry	Productive Systems

	machine readable form using standard W3C vocabularies is widely adopted.		

8.4.3 Big Data and Content Analytics

In the application fields mentioned above, processing large amounts of heterogeneous and multilingual data is a must. The challenge for the years to come consists in **developing robust techniques that can process and sift through large amounts of data to generate insights in near-real time.** As most of the content generated is still of an unstructured nature, **Big Data techniques need to be applied to efficiently process large amounts of unstructured data and combine it with existing structured data as well as data in other modalities** (audio, video, image, 3D data, etc.). Following the trend identified by Gartner to provide analytics services on the clouds, it seems reasonable to assume that these **solutions for processing unstructured content at large scale will be deployed and run on the cloud at the place where the data resides.** Currently, linguistic data provision and provision of analytics services is fragmented, with many services existing using different standards, data formats, licenses, etc. It is crucial to develop an **ecosystem of linguistic data and content analytics services where data and services operate together seamlessly**. Adoption of open web standards (e.g. RDF, SPARQL, OWL, JSON) as well as modern web service technology (e.g. REST) is key towards accomplishing this ecosystem. **Towards processing Big Data, workflows need to be distributed so that processing can be efficient.** For this, **deploying workflows and NLP services on the cloud needs to be facilitated, allowing people to easily configure a workflow and then executing the workflow on the cloud.** Standard interfaces and APIs needs to be provided and implemented by legacy and new services to make them easily integrateable and composable to address more complex tasks. In order to realize such an ecosystem of high-throughput services that implement compatible APIs and interfaces and are thus composable into complex workflows, the following is needed:

- **Standard APIs and interfaces:** Defining standard APIs and interfaces based on W3C and other standardized vocabularies defined by the linked data community (e.g. NIF, *lemon*, etc). This will allow to easily combine and exchange services easily and make the task of integrating them into existing workflows easier.
- **Best practices for publishing of linguistic linked data resources:** Best practices for publishing and sharing linguistic linked data resources are developed for the most important types of resources.
- **Best practices for implementation of NLP services:** Best practices for the implementation of content analytics services that are directly layered on top of the current web architecture are developed, requiring only HTTP as protocol but no additional protocols such as SOAP or other RPI methods.
- **Flexible LLOD-aware service architectures:** New LLOD-aware architectures for service deployment and composition that build on current Web standards such as JSON-LD are developed. This new generation of LLOD-aware services can consume LLOD resources and produce new resources that are dynamically added to the LLOD cloud.
- **Queryable repositories of linguistic linked data:** Well-known repositories and providers of linguistic resources expose metadata following Web standards such as SPARQL and standard W3C vocabularies to support querying repositories by machines.
- **Service deployment on the cloud:** Infrastructure is created to instantiate and deploy content analytics workflows on the cloud as well as to train NLP components with datasets as part of the LLOD and deploy these on the cloud.
- **Certification and validation of data and services:** Infrastructure for validating, benchmarking and certifying the performance and quality is available and implemented by independent parties, e.g. aggregators of services.

- **Multilinguality:** Tools that extract all the relevant entities (see Section 5) in all European languages need to be available.
- **Variety:** Approaches that can efficiently process different types of resources including text, structured as well as other modalities (video, audio, images) become available.
- **Scalability:** In order to scale to large amounts of content, content analytic services need to effectively rely on distribution and parallelization as well as optimally balance offline and online computation to support real-time performance.

Some of the above requirements were identified as part of the 2nd LIDER roadmapping workshop collocated with the Multilingual Web Workshop 2013 in Madrid (see here and LIDER deliverable D4.5)

Roadmap for Big Data and Multilingual Content Analytics

Horizon	Prediction	Actors	Means
1-2 Years	Standard interfaces for NLP services based on open standards and vocabularies are defined.	Academia-Industry Partnerships	Standardization
1-2 Years	Best practices for publishing linguistic resources fostering exploitation by content analytics services emerge.	Academia-Industry Partnerships	Research Projects and Pilots
1-2 Years	Proposals for flexible architecture for LLOD-aware services emerge.	Academia-Industry Partnerships	Research Projects and Pilots
3-5 Years	Infrastructure and standards for easy deployment of services on the cloud emerge and are increasingly used by content analytics providers and users.	Academia-Industry Partnerships	Standardization, Research Projects and Pilots
3-5 Years	New approaches for scalable content analytic services that are deployed on the cloud and rely on distribution and parallelization emerge and are increasingly used.	Industry	New Products and Services
3-5 Years	Standard interfaces and APIs for NLP and content analytic services are widely adopted.	Industry	Productive Systems
5-10 Years	Best practices and approaches for balancing online and offline computation in content analytics to provide answers in real time emerge and are applied.	Academia & Industry	Integrated Projects and Productive Systems
5-10 Years	Robust approaches to process heterogeneous data sources efficiently are developed and used.	Academia & Industry	Integrated Projects and Productive Systems
5-10 Years	A landscape of mature tools that support extraction of relevant entities (topics, named entities, facts, relations etc.) are available for all European languages and are used in industrial contexts.	Industry	Productive Systems

9. Conclusion

In this report we have analyzed the potential impact of linked data technology, in particular, linguistic linked data, on content analytics. We have identified a number of key application fields in which linked data technology can be expected to have a major impact, and we have identified the steps together with their time frames required to unleash this potential impact and provide a clear added value in the application areas identified.

In the application field of **Global Customer Engagement**, linked data has a strong potential to support the creation of rich multimedia experiences that deliver content to users as a multichannel and multimodal experience, supporting also storytelling. Linked data will

contribute to techniques for integrating and linking data across sites, media and languages and it will support repurposing of content with respect to language, modality and audience. It will support the semantic description of people, products, contexts and situations, and intentions and it will exploit these descriptions to provide tailored, personalized and contextualized user experiences and interactions. We foresee that current advertising and customer interaction models in which content or recommendations are pushed to customers will radically change towards models in which single customers and collections of customers are addressed via a (moderated) dialogue. This will require advanced techniques for robustly analyzing the voice of the customers including their intents and desires, for all European languages, but also techniques that automatically generate content in real time, also in multiple languages. Lexicalized and multilingual linked terminologies and ontologies will play a major role here to ensure consistency of messages as well as in supporting the repurposing of content across languages, modalities and audiences.

In the area of **Public Sector and Civil Society**, linked data will contribute to create an ecosystem of data that is partially open and partially closed but is extended with appropriate provenance and licensing information as well as mechanisms for representing and dealing with trust and confidence, so that the public as well as private companies can exploit the data for their purposes and within their applications. Simplifying access to data by appropriate interfaces, e.g. based on natural language, is a crucial goal to achieve. Linked data is a crucial part to create the digital commons and to distribute data and make it accessible via SPARQL views, thus fulfilling basic criteria to develop secure and trustful information networks. Most importantly, linguistic linked data technology has the potential to contribute to the creation of a single digital market in which services operate across the borders of languages and nations. For this, relevant application fields in which communication can be elevated to a language-independent semantic level need to be identified, and appropriate vocabularies need to be developed. Further, linked data can contribute to linking vocabularies, terminologies and ontologies from different linguistic and national contexts as a basis to achieve interoperability as well as as a basis for the localization of services.

With respect to the **Linguistic Linked Data Life Cycle and Linguistic Linked Data Value Chain**, linked data technology has the potential to effectively improve the way that linguistic data is published and shared, by making it easier to discover and exploit in current workflows and applications. For this, current providers of linguistic metadata should adopt semantic and linked data technologies to expose metadata in addition to provenance and licensing information so that discovery of relevant datasets becomes possible, also by machines. Adoption of standardized vocabularies for metadata description but also for datasets (e.g. *lemon*) is crucial to support interoperability. In order to cater for the large need for linguistic resources, a language resource market that builds on linked data needs to emerge, creating new models of revenue that include non-monetary benefits for those who create, curate, refine or extend existing resources. Call for tenders for the creation of linguistic resources could turn to be a flexible and efficient mechanism towards this end. Collaborative methodologies for the creation of linguistic resources in which each actor is recompensed in some way for their contribution will need to emerge.

Building the above sketched ecosystem for linguistic linked data requires a landscape of LLOD-aware natural language processing and content analytics services that exploit linguistic linked data effectively and, in addition, i) are discoverable, ii) rely on standardized APIs and interfaces, iii) are easily integrateable into standard workflows as well as composable, and are iv) scalable. Scalability is particularly important to make this ecosystem of LLOD-aware services ready to support Big Data applications in which large amounts of unstructured data (including texts, videos, images but also other modalities) need to be analyzed efficiently. Building on distribution of services over the cloud and data as well as supporting stream scenarios is crucial in this context.

A crucial issue for the future is to appropriately deal with intellectual property, copyright but also privacy. In all these fields we foresee a strong potential for linked data technology as a way to ensure that provenance and licensing information remains attached to the data over the whole lifecycle, thus supporting awareness of the copyright and provenance of data, also

when it is aggregated and mashed up with other data, as well as to monitor compliance with terms and conditions. Linked data vocabularies as well as the RDF language represent and effective means to make such information machine-readable and will play a crucial role in supporting machine-to-machine negotiation.

In order to boost adoption by SMEs, the cost-effectiveness of solutions is critical. There also need to be clear guidelines for publishing and consuming data sources as linked open data. Proof-of-concept implementations and show cases can help to analyze benefits and cost for a larger audience, in joint partnerships between academia and industry.

10. References

1 Robert Pepper and John Garrity. The internet of everything: How the network unleashes the benefits of big data. In The Global Information Technology Report 2014. World Economic Forum, 2014.
2 M. Palmer. Data is the new oil, November 2006,
3 C.P. Alexander. The new economy. Time Magazine, May 1983.
4 SINTEF. Big data, for better or worse: 90% of world's data generated over last two years. ScienceDaily, May 2013
5 J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, December 2012
6 Beñat Bilbao-Osorio and Soumitra Dutta and Bruno Lanvin (eds.), The Global Information Technology Report, World Economic Forum, 2014
7 Anjali Yakkundi, David Aponovich, and Mark Grannan. TechradarTM For AD&D Pros: Digital Customer Experience Technologies, QS 2013. Forrester Research, 2013.
8 Alex Pentland. Big data: Balancing the risks and rewards of data-driven public policy. In The Global Information Technology Report 2014. World Economic Forum, 2014.
9 Marc Teerlink, Paula Wiles Sigmon, Brett Gow and Kingshuk Banerjee. The new hero of big data and analytics: The Chief Data Officer, IBM Global Business Services, Executive Report, 2014.
10 Mark Cieliebak, Oliver Dürr, and Fatih Uzdilli. Potential and Limitations of Commercial Sentiment Detection Tools. In: Proc. of the Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI, 2013
11 LT Innovate. The LT-Innovate Innovation Manifesto. LT Innovate, 2014.
12 A.T. Kearney (ed.). Rethinking Personal Data: A New Lens for Strengthening Trust. World Economic Forum, 2014.
13 European Commission. Connecting Europe Facility: Investing in europe's growth, September 2012
14 Komninos, Nicos.Intelligent. Cities: towards interactive and global innovation environments. International Journal of Innovation and Regional Development (Inderscience Publishers) 1 (4): 337–355(19), 2009, doi:10.1504/ijird.2009.022726
15 Meta Technology Council. Strategic Research Agenda for Multilingual Europe 2020. Springer, 2012.
16 LIDER project. [https://www.w3.org/community/ld4lt/wiki/Lider_roadmapping_activities Lider roadmapping activities], 2013
17 European Commission 2012, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), available at http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf
18 Cisco (2013), "Connections Counter: The Internet of Everything in Motion." In: the network: Cisco's Technology News Site, July 29. available at http://newsroom.cisco.com/feature-content?type=webcontent&articleId=1208342