LIDER: FP7 – 610782

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

Deliverable number	**D1.1.1**
Deliverable title	**Business use cases for the use of Linguistic Linked Data in content analytics processes - Phase I**
Main Authors	**Kevin Koidl, David Lewis, Paul Buitelaar**

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	D1.1.1
Class Deliverable	LIDER EU-ICT-2013-610782
Version	0.1
Document due date	30 April 2014
Submitted	<Day Month Year>
Responsible	Kevin Koidl (TCD)
Reply to	koidlk@scss.tcd.ie
Document status	initial draft
Nature	R(Report)
Dissemination level	PU(Public)
WP/Task responsible(s)	WP1 Business use cases/Task1.1 Collection of business use cases for the use of Linguistic Linked Data in content analytics processes
Contributors	Kevin Koidl (TCD), David Lewis (TCD), Paul Buitelaar (NUIG)
Distribution List	Consortium Partners
Reviewers	Felix Sasaki, ERCIM
Document Location	http://www.lider-project.eu/?q=doc/deliverables

Executive Summary

This deliverable presents the initial results of the effort in the LIDER project related to WP1. The goal is to understand the current and potential industrial needs in content analytics processes regarding Linguistic Linked Data. For this a survey and its results are presented.

In addition project partners and institutions that are part of the Industrial Board have engaged to identify the common content analytics tasks and use cases in their current work practices and the possible benefits from the use of Linguistic Linked Data. This resulted in seed use cases reflecting existing work and also by requirement and use case prioritization capture from participants at the Linked Data for Language Technology Roadmapping Workshop (21st of March in Athens, Greece).

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://www.lider-project.eu/?q=doc/deliverables		
EU Project Officer	Susan Fraser		

Deliverable	Number	D1.1.1	Title	Business use cases for the use of Linguistic Linked Data in content analytics processes - Phase I
Workpackage	Number	WP1	Title	Business use cases

Date of Delivery	Contractual	30 April 2014	Actual	30 April 2014
Status	Version 1		Final	
Nature	Prototype, report, dissemination			
Dissemination level	Public, consortium			

Authors (Partner)	TCD			
Responsible Author	Name	Kevin Koidl	E-mail	kevin.koidl@scss.tcd.ie
	Partner	TCD	Phone	000353- (0)1 - 8963466

Abstract (for dissemination)	This deliverable presents the initial results of the effort in the LIDER project (WP1) to understand the current and potential industrial needs in content analytics processes regarding Linguistic Linked Data.
Keywords	use cases, business cases, survey

Version	Modification(s)	Date	Author(s)
01	Document Draft	24/4/2014	Dave Lewis (TCD), Kevin Koidl (TCD)
02	Document Internal Review	25/4/2014	Felix Sasaki (ERCIM)
03	Document Final Corrections	28/4/2014	Kevin Koidl (TCD)

Project Consortium Information

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

Table of Contents

1 INTRODUCTION ... 7
1.1 METHODOLOGY ... 7
2 STATE OF THE ART: SEED USE CASES ... 8
2.1 ONTOLOGY LOCALISATION ... 8
2.2 PUBLISHING RICH LEXICAL KNOWLEDGE WITH ONTOLOGIES ... 9
2.3 AGGREGATION OF LEXICAL AND ENCYCLOPAEDIC SOURCES ... 10
2.4 MULTILINGUAL DISAMBIGUATION AND ENTITY LINKING ... 10
2.5 MULTILINGUAL AND CROSS-LINGUAL SENTIMENT ANALYSIS ... 11
2.6 TERMINOLOGY EXTRACTION ... 12
2.7 SPEEDING UP GRAMMAR GENERATION FOR SPOKEN DIALOGUE SYSTEMS ... 12
3 SURVEY ANALYSIS ... 13
3.1 PARTICIPANT PROFILE ... 13
3.2 NLP APPLICATION AREAS ... 14
3.2.1 Discovering and Extracting Information ... 15
3.2.2 Understanding Opinion ... 15
3.2.3 Data Management ... 16
3.2.4 Monitoring and Forecasting ... 16
3.3 USE OF LANGUAGE RESOURCES ... 17
3.4 AWARENESS/MATURITY IN USING LINKED DATA ... 17
4 ROADMAPPING WORKSHOP RESULTS ... 18
5 CONSOLIDATED USE CASES PROVIDED BY THE COMMUNITY ... 20
6 REVIEW OF APPROACH ... 21

1 Introduction

This deliverable presents the initial results of the effort in the LIDER project (WP1) to understand the current and potential industrial needs in content analytics processes regarding Linguistic Linked Data. This is advanced using the community engagement channels organized by WP4 and described in deliverable D4.1 (community building and dissemination Plan) and reported on in D4.3.1 (Preliminary report on LIDER Community and community portal – Phase I)

Members of the Industrial Board (constituted in WP4) have been engaged to define a set of business use cases to describe the use of Linguistic Linked Data in content analytics processes (Task 1.1). The consortium, with the assistance of some members of the Industrial Board has conducted an initial analysis of these use cases to extract those requirements needed to exploit Linguistic Linked Data in content analytics processes and to identify the common and frequent tasks in content analytics that require NLP and Linguistic Linked Data (Task 1.2). This report (D1.1.1) offers initial results of this process and a second a final report D1.1.2) resulting from further informed engagement will be published in April 2015.

These reports will provide phased input for WP2, to produce a set of guidelines and identify best practices, and into WP3 to prepare a roadmap for the use of Linguistic Linked Data in content analytics processes.

1.1 *Methodology*

Project partners and institutions that are part of the Industrial Board have engaged to identify the common content analytics tasks in their current work practices and the possible benefits from the use of Linguistic Linked Data. The identification of these tasks relied on formation of the Industrial Board by T4.1, in the form of the Linked Data for Language Technology (LD4LT) W3C Community Group[1].

To captures previous work on use cases for linguistic linked data in order to inform face to face and group discussion, these use cases were assembled and summarized in section 2.

An initial online questionnaire has be deployed via the LD4LT community Group. This elicited information on language technology application areas of interest, the levels of awareness/maturity in using linked data and their industry sectors. This is reported in section 3.

Based on the uptake and outcome of this questionnaire, guidelines were developed to enable the WP1 team to drill down with Board members to elicit further details. This was done through a mix of structured interviews and group discussion which identified business pain points resulting from missing or siloed meta-data or poor access to text/media analytics, or success stories where open meta-data and analytics played an important role. These were conducted through a combination of telephone interview and

[1] www.w3.org/community/**ld4lt**/

group discussion at roadmapping workshops organized by WP4. For this initial report the the group discussion were undertaken at a roadmapping workshop co-located with the European Data Forum, in Athens on the 21st March 2014. The outcome of this workshop is reported in section 4.

These inputs were collected and collated for analysis by T1.2 to produce an initial set prioritised and structured use cases defined in section 5. This will provide input to the best practice, reference architecture and research roadmapping exercises in LIDER on how Linguistic Linked Data and those NLP services built on top of them have or could have solved concrete business issues.

Subsequent to the publication of this deliverable, the process will be rerun based on the critical review of the process presented in section 6. This will involved specific guidance on presenting to Board members possible solutions resulting from best practice and architecture work from WP2 and WP3 as well as working with additional industrial contacts developed through WP4. This will provide feedback on proposals in terms of technical, legal or social barriers that enterprises may identify as well as providing input into a more refined and viable set of use cases (presented as part of D1.1.2) to guide further best practice and architecture work.

2 State of the Art: Seed Use Cases

This is taken from https://www.w3.org/community/ld4lt/wiki/Seed_Use_Cases and is continued to be used for further dissemination.

2.1 *Ontology Localisation*

Title: Ontology Localization

Source Reference: http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html

Industry sector: Any

Actors and benefits they get from use case: Developers of ontologies, controlled vocabularies that want to localize their ontologies for cross-lingual interoperability

Summary: In many scenarios, the localization of ontologies is an important task as interoperability needs to be established across languages and borders. The Monnet project has considered in particular the case of financial terminologies / vocabularies such as XBRL and the specific GAAP-based language-specific vocabularies. In order to establish interoperability between all these national vocabularies, they need to be aligned. As a basis for automatic or semi-automatic alignment, appropriate translations are needed. Thus, techniques are needed to translate the labels of an ontology into some other target language, which we refer to as ``ontology localization.

Language technologies involved:
Automated terminology extraction and analysis machine translation systems

Language resources involved: Multilingual or monolingual linked data lexicon or dictionary multilingual term bases translation memories

Specific benefit of using linguistic linked data in use case:
Reuse resources for finding (domain-specific) translation candidates supporting the localization

Provided by: Philipp Cimiano - UNIBI

2.2 *Publishing Rich Lexical Knowledge with Ontologies*

Title: Publishing Rich Lexical Knowledge with Ontologies

Source Reference:
http://www.w3.org/community/ontolex/wiki/IFLA, http://www.w3.org/community/ontolex/wiki/AGROVOC

Industry sector: Localisation, specifically terminology management

Actors and benefits they get from use case: Developers of thesauri, classification schemes, ontologies Consumers and users of thesauri, classification schemes,

Summary: In many cases a deeper linguistic grounding of linguistic elements specified in thesauri is required, e.g. including inflectional or other syntactic information describing how the terms behave syntactically and semantically. However, current models such as SKOS are not sufficient for this purpose. Thus, there is a clear need for an extensive vocabulary to model the linguistic properties of terms in a thesaurus, classification scheme etc. This need has been documented at http://www.w3.org/community/ontolex/wiki/IFLA for the International Federation of Library Associations and Institutions and athttp://www.w3.org/community/ontolex/wiki/AGROVOC for the AGROVOC thesaurus developed by the Food and Agriculture Organization (FAO)
Several technological needs arise in the context of such a use case, such as i) easy porting of SKOS resources to ontolex, automatic creation of ontolex resources from non-ontolex resources (SKOS, RDF etc.). The needs of the above example use cases is addressed by introducing the ontolex vocabulary currently in development by the ontolex W3C community group, which provides the vocabulary necessary to define ontology lexica that realize a separate lexical layer that can be used to provide the rich linguistic and lexical information externally to the ontology / thesaurus / classification scheme in question in a separate file
The benefit of the ontoloex model is that ontology lexica can be published separately from the ontologies / models they lexicalize, thus giving a high degree of flexibility to add lexica for additional languages. This in particular adds some modularity as support for other languages can be added incrementally as needed.

Language technologies involved: Morphological analysers translation tools

Language resources involved: Multilingual or monolingual dictionaries
existing lexical resources to link to.

Specific benefit of using linguistic linked data in use case: Reuse of available resources for localization of lexica.

Provided by: Philipp Cimiano, UNIBI

2.3 *Aggregation of Lexical and Encyclopaedic Sources*

Title: Aggregation of Lexical and Encyclopaedic sources
Source Reference: http://multijedi.org http://babelnet.org http://babelnet.org/2.0
Industry sector: Any
Actors and benefits they get from use case: Developers of dictionaries, encyclopedias, thesauri, ontologies. Consumers and users of large machine-readable, language knowledge resources (e.g. companies building systems that require large amounts of knowledge)

Summary: The alignment and integration of lexicographic, i.e. from dictionaries, and encyclopedic knowledge, i.e. from encyclopedias, is crucial for both developers and consumers of large knowledge resources. However, many online language resources are either based on Wikipedia, e.g. DBpedia, thereby focusing on encyclopedic content, or on dictionaries, such as OmegaWiki or Wiktionary. The MultiJEDI project has considered the case of interlinking and merging several language resources, i.e., WordNet, Wikipedia, OmegaWiki and the Open Multilingual WordNets. To perform the alignment techniques are needed which link the same meanings available in different resources and decide when to merge the corresponding concepts into unified, multilingual concept representations.

Language technologies involved: Word sense disambiguation machine translation tools

Language resources involved: Multilingual or monolingual online dictionaries and encyclopedias

Specific benefit of using linguistic linked data in use case: Reuse of available resources Alignment to and exploitation and availability of other language resources in the LLOD cloud

Provided by: Roberto Navigli, UNIRM

2.4 *Multilingual Disambiguation and Entity Linking*

Title: Multilingual Disambiguation and Entity Linking

Source Reference: http://babelfy.org (available online by the end of April)

Industry sector: Any

Actors and benefits they get from use case: Users and consumers of semantically-annotated or semantically-indexed text/data in any language

Summary: While Word Sense Disambiguation, the task of automatically associating meanings with words in context, has typically been a task restricted to a small number of

researchers, recently the emerge of the new task of Entity Linking, concerned with linking named entities within text, has opened up new possibilities for a huge number of companies in search for services aimed at semantic indexing and linking of text written in arbitrary languages.
To perform Entity Linking, however, large amounts of machine-readable knowledge need to be available, together with effective algorithms for performing the task.
While several approaches to Entity Linking exist, the MultiJEDI project has addressed the task of integrating Word Sense Disambiguation with Entity Linking, showing that Wikification and Entity Linking services can greatly benefit from the integration of lexical, i.e. from dictionaries, and encyclopedic knowledge.
To do this and keep the task independent of language, large amounts of knowledge, lexicalized and connected in as many languages as possible, need to be made available.

Language technologies involved: Word sense disambiguation entity linking

Language resources involved: BabelNet, multilingual or monolingual online dictionaries and encyclopedias

Specific benefit of using linguistic linked data in use case: Performance improvement thanks to linking to and exploiting other LLOD

Provided by: Roberto Navigli, UNIRM

2.5 *Multilingual and Cross-lingual Sentiment Analysis*

Title: Multilingual and Cross-lingual Sentiment Analysis

Source Reference: Slides of WebLyzard at EDF LIDER WS Meeting

Industry sector: Any

Actors and benefits they get from use case: Developers of sentiment analysis / opinion mining systems Sentiment analysis and opinion mining systems are heavily used to understand and structure online communication about products, services etc. for marketing purposes. In many cases, analysis of brands across countries and natural languages is crucial. However, adopting a sentiment analysis system to other domains is expensive, requiring sentiment lexica in different languages, which are ideally contextualized.

Language technologies involved: Automated sentiment analysis

Language resources involved: Contextualized sentiment lexica for multiple languages

Specific benefit of using linguistic linked data in use case: Reuse (linked) sentiment lexica in multiple languages to adopt a sentiment analysis system to other languages, lowering the cost for doing so

Provided by: Philipp Cimiano - UNIBI

2.6 *Terminology Extraction*

Title: Terminology Extraction for Localisation

Source Reference: FALCON Project

Industry sector: Localisation, specifically terminology management

Actors and benefits they get from use case: Localisation clients who will improve the terminology consistency of source content prior to translation Language Service Providers who will be able to provide better translation quality through consistency in term translation.

Summary: Localisation client runs source text through an automated term identification service which has been trained on a dictionary indexed with references to one or more linked data dictionaries, including their own organisation one
The automated term identification service returns the source text with terms annotated, e.g. using ITS2.0 terminology data category, with a reference to a lexical data entry.
The term annotations are reviewed by the client's terminologist who identifies any false positives, dereferencing and reviewing information in the lexical data entry if needed. False positives may be fed back to improve the training corpora of the terminology extraction service.
The approved term annotations are then reviewed by a linguist who may, if the dictionary is multilingual and includes the source language, then deference the link of the terms, examine any translations present and opt to approve it for use in the job. If a translation is not present the linguistic may provide one if judge to be important for translation consistency. In both cases the term translations are then passed as a multilingual glossary together with the source text to the language service provider. If a new translation had been generated, this may be submitted back to the dictionary as a candidate translation for future use.

Language technologies involved: Automated terminology extraction
term suggestion and translation review tool.

Language resources involved: Multilingual or monolingual linked data lexicon or dictionary.

Specific benefit of using linguistic linked data in use case: Critical assessment of use case, e.g. in commercial use, trial only, research prototype, under-development

Provided by: Dave Lewis - CNGL/TCD

2.7 *Speeding Up Grammar Generation for Spoken Dialogue Systems*

Title: Supporting Development of Dialogue Systems using Linked Data and Ontology Lexica

Source Reference: Portdial Project

Industry sector: Any

Actors and benefits they get from use case: Developers of dialogue systems

Summary: The creation of grammars for dialogue systems is costly. It can be made more cost-efficient by techniques that semi-automatically support the creation of grammars. In the Portdial project, an approach by which grammars are induced in a top-down fashion from an ontology lexicon has been explored and shown to deliver grammars that are highly precise but lack recall. UNIBI has implemented this approach for LTAG (Lexicalized Tree Adjoining Grammars), CCG (Combinatorial Categorial Grammars) and GF (Grammatical Framework Grammars)
Nevertheless, such grammars can be used to seed the further process of extending the coverage of the grammar, using other techniques to increase coverage. In addition, the Portdial has in particularly supported the development of so called pre-terminal rules which expend non-terminals into a set of named entities. It has been shown that Linked data can be used to support this use case.

Language technologies involved: Top-down grammar generation

Language resources involved: Linked data with labels in different languages
existing language-specific grammars

Specific benefit of using linguistic linked data in use case: Reuse (linked) ontology lexica for top-down grammar induction reuse linked data with labels in many languages to support enhancement of pre-terminal rules in many languages.

Provided by: Philipp Cimiano - UNIBI

3 Survey Analysis

The main goal of this survey is to understand current industrial needs, requirements and use cases that will help define a roadmap for future R&D activities in multilingual/multimedia content analytics. Other implicit goals of this survey were to improve awareness of the potential of linked data for NLP applications as well as to make known existing expertise in this area in Europe. Another implicit goal was to identify potential partners for research. The survey was available online, recruiting participants by email using our contacts as well as on different mailing lists.

A total of 27 participants were recruited for this survey. The questions considered in our survey are organized in four main parts. The parts are the following: participant profile, NLP application areas, use of language resources, and awareness/maturity in using linked data.

3.1 *Participant Profile*

The first part of the survey is concerned with gathering information about the profile of each participant. Participants were asked about the type of organization they are associated with and the industry sector they are active in, allowing them to choose between multiple options. While circulating this survey, we specifically stated our interest in industry participation. Therefore the majority of responders are affiliated with SMEs or large organizations, and only a smaller number with public sector organizations, as can be seen in Table 1. Each participant can have more than one affiliation and can be active in multiple industry sectors. Therefore, participants were allowed to choose more than one option in both cases.

Organization type	Responders
SME	13
Large Organization	6
Public Sector	6
Non-profit	1
Freelancer	0
Other	1

Table 1. Organisation type

The second question about participant profile was related to industry sectors. A list of industry sectors was provided to the participants, but they were also given the option to choose a miscellaneous category, called “Others”. Table 2 presents their responses, showing a marked interest from professionals in Public Sector Publishing, Media, News and Journalism and Localization. Other sectors that showed interest in the area of using linked data in NLP applications for content analytics included the Pharmaceutical sector, Service/Product vendors and eHealth.

Industry sector	Responders
Other	10
Public Sector publishers	9
Media, News and Journalism	8
Localization	7
Pharmaceutical	6
Service / Product vendors (customer support)	6
eHealth	5
Content Management Tool Vendors	3
Libraries, Museums, Digital Humanities	3
Finance	2
ePublishing / eBook	2
eEnergy	1
eTransport	1
Peer production communities	0

Table 2. Industry sector

3.2 *NLP Application Areas*

The second part of the survey is concerned with identifying NLP applications in content analytics that are of interest to the industrial community. Several broad areas of applications were identified, including Discovering and Extracting Information, Understanding Opinion, Data Management, and Monitoring and Forecasting. Overall, the use case that achieved the highest consensus with respect to industrial interest is related to the extraction of information from unstructured data, from the broad area of Discovering and Extracting Information.

The following sections discuss industrial interest in each of these areas separately.

3.2.1 Discovering and Extracting Information

The first application area brings together several use cases related to information discovery and information extraction, which are listed in Table 3. The majority of the responders identified the area of information extraction from unstructured data as an area they are interested in. Other areas that gathered a majority of votes included entity and event detection, expert finding and semantic search.

Use case	Responders
Extraction of information from unstructured data	24
Entity and event detection	18
Expert finding from unstructured and structured data	18
Semantic search	18
Text-to-semantics conversion	11
Question answering in natural language	10
Multimedia and video search, visual search	9
Fact validation using unstructured / web data	8
Speech-to-semantics conversion	5

Table 3. Use cases related to discovering and extracting Information

3.2.2 Understanding Opinion

The second application area groups together use cases related to the broad area of understanding opinions. An overview of the responses in this area is presented in Table 4. A large number of participants identified impact analysis as a relevant use case, immediately followed by use cases in sentiment and opinion mining. Other use cases of interest for the industrial community include mining customer interaction data and trend mining, with almost half of the participants expressing interest in them.

Use case	Responders
Impact analysis (e.g. of marketing campaigns or other marketing measures)	16
Sentiment / opinion mining	15
Mining customer interaction data to acquire insights about their behavior	13
Trend mining	13
Identifying key opinion holders / opinion leaders	12
Identifying and making explicit the argument structure and logical relation between opinions within public discourse about a topic	11
Identifying (potentially) opposing communities	5
Identifying irony / sarcasm in web texts / reviews	5

Table 4. Use cases related to understanding opinion

3.2.3 Data Management

The area of data management organizes several NLP application areas related to creating, organizing, sharing and storing content and data, which are listed in Table 5. The largest number of participants showed an interest in use cases related to data integration and content summarization. More than half of the participants expressed an interest in tools that support ontology building, evolution, and maintenance and topic detection.

Use case	**Responders**
Data integration	17
Content (text, multimedia) summarization	15
Support for text-based ontology building / evolution / maintenance	14
Topic detection	14
Aspect oriented data summarization	13
Rapid knowledge base formation from textual data for analytics task	13
Supporting development of (multilingual) terminologies / thesauri / term bases	13
Taxonomy maintenance	11
Machine translation	8
Speech-to-text conversion	8
Natural language generation from templates, database content etc.	7
Digital preservation of multilingual, multimedia content	6
Information kiosk	6
Multimedia eLearning	6
Speech processing	4
Computer and video games	1

Table 5. Use cases related to understanding opinion

3.2.4 Monitoring and Forecasting

The last broad area considered in this survey is related to monitoring and forecasting topics and entities of interest, described in more detail in Table 6. The most relevant use case to the industrial community was the use case related to predictive analytics over text data, followed by use cases that address tracking entities on the Web.

Use case	**Responders**
Predictive analytics over text data	18
Tracking entities (people, products) on the Web	15
What-if-simulation based on content analytics results finding relevant communities/fora/discussion pages on the Web	6

Table 6. Use cases related to understanding opinion

3.3 *Use of Language Resources*

This part of the survey was concerned with mapping industrial use of existing language resources. Participants were asked about the type of language resource that they make use of in their daily activities, as can be seen in Table 7. Dictionaries, corpora, and tokenizers are the most widely used resources by the industrial community.

Language resource	Responders
Dictionaries (Monolingual / Bilingual / Multilingual)	15
Corpora (Written / Spoken / Multimodal)	13
Tokenizers	12
NLP Frameworks: UIMA / GATE / NLTK Toolkit	11
Part-of-speech Taggers	11
Sentence Splitters	11
Terminologies	11
Encyclopedic resources (DBpedia, YAGO, BabelNet, etc.)	9
Parsers	9
Term bases	9
Translation memories/parallel text	9
Machine Translation Systems (e.g. Moses, Google, Bing, …)	8
Others	5

Table 7. Type of language resource

The second question related to the use of language resources was concerned with the location of the language resource used. The majority of the participants make use of a mixture of language resources that are produced within their organization together with external language resources, as can be seen in Table 8.

Language resource location	Responders
In-house	6
External language resources	4
Both of the above	17

Table 8. Location of language resource

3.4 *Awareness/maturity in using Linked Data*

The last part of the survey gathers information about the awareness and maturity of using Linked Data and Linguistic Linked Data, in Tables 8 and 9, respectively. Not surprisingly, the majority of the participants are very aware of Linked Data.

Awareness	Responders
Very aware	16
Not so	8
Not at all	3

Table 9. Linked Data usage

The same thing cannot be said about Linguistic Linked Data, because less than half of the participants stated that they are aware of this resource.

Awareness	Responders
Very aware	12
Not so	9
Not at all	6

Table 10. Linguistic Linked Data usage

4 Roadmapping Workshop Results

The first LIDER industry board roadmapping workshop was help on 21st March 2014 in Athens. A full report will be made available as deliverable D4.5 “Report of the 1st Roadmapping Workshop”, due end of May 2014. Here we just summarise the immediate workshop results that informed the consolidated set of use cases reported in section 5.

During the road mapping workshop the participants were asked to indicate uses cases and related requirements on a flip chart.

Following use cases were communicated by the participants:

	Participant Use Cases
1	Align entities from different language resources
2	Disambiguation by sense
3	Ontology for support my LR Life Cycle
4	Linkage to non LR catalogues, to foster new usage scenarios
5	Handling of inflected forms / variables / accentuation in automated way
6	Use Existing Formats
7	Conversions to Industry Formats TMX, TBX, etc.
8	Open (documental) standards
9	Combination of heterogeneous data packages & automatic metadata description
10	Compatibility with already existing and widely used cat-tools (Computer-Aided translation)
11	Tools agnostic
12	Sharing resources
13	Unique "Ontology" metadata schema for lang. resources
14	Community agreement on shared vocabularies
15	Easy identification of LRs documentation etc. in an uniform way
16	Should be domain agnostic
17	Model usage of data
18	Find Business Models For Sharing Data
19	Pricing Models & Commercial Transactions
20	Open Source vs. Commercial
21	Crowdsourcing curation
22	License Retrieval, Management And Control
23	Access Control-Pay-Portal

24	Licensing
25	IPR
26	Third Party Quality Annotation
27	Multilingual Dictionaries Dynamic Creation/Aggregation
28	Document (e.g., news) Similarity Using heterogeneous L.D.
29	Brand Monitoring
30	Sentiment-Based Recommendations
31	Sentiment Analysis
32	Organizational / corporate knowledge graphs
33	Build a knowledge Interlinked Graph in various languages
34	Organic. Edunet Multilingual discovery service (How models like train the trainer)
35	Trend Analysis
36	Machine Translation: Need to disambiguate senses on source side
37	Cross lingual data discovery/Navigation
38	Use of NLP services and L resources for data cross lingual interlinking
39	Cross-lingual and/or Cross-format resource linking
40	Accessibility and availability of language data
41	Parallel (MT) Data Discovery
42	Translation Industry
43	Global Translation Memory
44	Modeling Life cycle of LRs recording as LOD all the steps
45	Ontological resources. Linked to dynamic data
46	Live localization: Use recurring concepts to make multilingual sites on-the-fly (e.g., bookings, recipes)

Following requirements were communicated by the participants:

	Participant Requirements
1	Minimal Core Metadata Vocabulary (RDFS) For Lang Resources
2	Common Owl Vocabulary to Represent Language Resources Metadata
3	Simple Service to Validate Metadata And Current Status Of LR
4	Preserving of Information available in existing meta-data
5	Integration with non RDF tooling
6	Promotion of impactful sets Of LRs in LD
7	Easily accessible good quality bilingual data by category list
8	Data validated by humans
9	QA precision: how to improve the quality of automatically generated LD
10	Means for collective Assessment
11	Model quality of language resources
12	Good entity extraction tools
13	Linguistic hub repository
14	Extensibility (versioning) How easy to add my additions provide "extension packages to existing resources"
15	Scalability (it should be fast) Distributed (local caches)
16	Services for discovery

17	New business models
18	One or two click button for migration LR into LD
19	Demonstrations/Tutorials of how resources can be used

5 Consolidated Use Cases provided by the community

The following industry focused use case was provided by industry.

Providing government services in Switzerland is complicated. It is a multilingual country, with three major languages (German, French, Italian), one minority language (Romansh), but also with a population that is 25% immigrants from around Europe and the world. The country has 26 cantons, and each of those cantons is responsible for its own services - health care, motor vehicles, education, etc. Four of those cantons are officially bilingual or multilingual, and provide government services in two or more languages. The remainder, containing about 4/5 of the population, provide services only in their single official language.

Official forms, such as tax documents or educational records, are always in the official language of the canton. Someone who lives in a French canton but works in the German section fills out all her paperwork and receives all her statements in German. Immigrants, who interact with the government regularly for work and residency permits, must muddle through in a language for which they are likely to not know bureaucratic or technical terms.

Yet, the information that governmental services need in order to interact with their people, which is collected on official forms, is strikingly consistent from one canton to the next. A child tax credit is the same thing in Zurich and Geneva, and date of birth does not change for someone who lives in French-speaking Jura and works a few kilometers away in German-speaking Basel.

These recurring concepts can be enumerated, translated, and served from a dictionary database to government agencies and their citizens. With the multilingual architecture we have developed at Kamusi, we can harvest terms that are used on official forms, and treat those as lexical items that can be defined in their original language and translated and defined in any other. Then an agency in German-speaking Lucerne could quickly print or produce on the web a French or Italian or Romansh version of, for example, the paperwork required to open a newsstand, building on concepts that have already entered the system via a sister agency in French-speaking Vaud. Moreover, the terms can be gathered in languages spoken by immigrant communities, whether English or Albanian or Portuguese, greatly reducing communications difficulties between the government and a quarter of the population.

This use case can be expanded in two directions:
1) Private companies and non-governmental organizations within Switzerland can use the same technology to build sets of translation terms for their forms and publications. Insurance companies and banks, for example, have predictable recurring concepts that the Kamusi system can easily produce as multilingual data. Medical records used by doctors and hospitals would require a greater amount of technical knowledge by each language's translators, but would also have a large payoff, for example in the ability to gather any patient's medical history when they arrive at a hospital.

2) Governments throughout Europe face the same challenges of dealing with multilingual communities, and have essentially the same set of recurring concepts on their forms. Data generated to facilitate interactions with the Spanish-speaking community in Switzerland would have similar use value with the Spanish-speaking community in Ireland, and the chain of links across languages means that the same concepts would be available for Romanians working in Spain. Once the data is produced in a language, it is available to governments throughout the continent - in essence, any agency in Europe could produce a form that, at the touch of a button, could be printed in any language.

6 Review of Approach

The survey and workshop was successful in seeding and exploring business focused use cases. Further work is required in following up feedback in one to one interview conversations.