FP7-ICT-2011.4.2
Contract no.: 288342
www.xlike.org

Deliverable D8.2.3

X-Like Showcase

Editor:	Esteban García-Cuesta, iSOCO
Author(s):	Esteban García-Cuesta, iSOCO; Gregor Leban, JSI; Marko Tadić, Božo Bekavac, UZG; Aljoša Rehar, STA.
Deliverable Nature:	Others (O)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M24
Actual Delivery Date:	M24
Suggested Readers:	All partners of the XLike project consortium and especially industrial partners and end-users, and EC.
Version:	1.0
Keywords:	Industry, demo, cross-lingual, dissemination, communication

Disclaimer

Full Project Title:	Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP8 – Dissemination, exploitation, and community building
Document Title:	D8.2.3 – XLike Industrial Showcase
Editor (Name, Affiliation)	Esteban Garcia-Cuesta iSOCO
Work package Leader (Name, affiliation)	Marko Tadić, UZG
Estimation of PM spent on the deliverable:	3 PM

Executive Summary

This deliverable is the last of two where the initial specifications for the showcase where provided in D8.2.1 (D8.2.1 XLike showcase specification) and this document shows the final fully functional demo ready for dissemination, commercial awareness and use by the European Commission and other targeted users.

It presents the fully functional XLike showcase associated to the event registry functionality which was introduced in showcase industrial specification which was presented at M6 in D8.2.1. This showcase makes use of different subsets of software and functionalities developed throughout the project. It also contains promotional material and dissemination plans for this showcase. The dissemination activities includes specific awareness activities such as industrial conference publications, video publication of the showcase in a real use case example, and full accessibility to the showcase in order to be used by the Commission or by any other organization which would like to make use of the XLike functionalities beyond the end of the project. The document also contains the description of the promotional material that has been created for the dissemination of the XLike project and more specifically for industrial awareness.

This document is related to and should be read together with D1.2.2, D4.3.1, D6.2.2, and D8.2.1.

Table of Contents

List of Figures

List of Tables

Abbreviations

REST	**Re**presentational **S**tate **T**ransfer
SaaS	**S**oftware **a**s **a** **S**ervice
SOA	**S**ervice **O**riented **A**rchitecture
T	**T**ask
XLike	**Cross**-**li**ngual **K**nowledge **E**xtraction
WP	**W**ork **P**ackage

1 Introduction

Nowadays, the journalist assistance for early detection of events is one of key elements of news production process due to the high competition between the different news media. The XLike project offers help in this task by providing a set of tools and functionalities for social media monitoring. These tools introduce an automatic way of harvesting, structuring, and accessing events which are occurring and which are being reported on in different information streams.

The presented showcase is an update of the [5] and also a concrete validation scenario for the cross-lingual event detection functionality. This showcase makes use of the complete set of functionalities implemented so far in the project as it is specified in [2]. This functionality is part of the WP4-WP5 (application layers) work packages and it is supported by the preceding ones WP1-WP3 (linguistic and semantic analysis).

The showcase scenario introduced at Section 2 aims to provide a real life environment where the generated technology can be deployed for helping actual end-users in a news agency (in our case STA). The description of the requirements specified at [5], needed to successfully accomplish the showcase scenario, is also described in this section including the relation between the different steps of the showcase scenario and those requirements.

The section 3 explains the different software/hardware components that have been used in the implementation of the showcase which can be split into two main parts: i) the architecture including the harvesting platform and analysis capabilities, and ii) the visualization which provides access to the different needed functionalities of the showcase scenario. In the current stage of development the platform allows the work almost in real time and we are able to fully analyse two thirds of the collected data. This percentage is good enough for the purpose of putting it all together and provides real capabilities and knowledge to the STA news agency. However, it does not deal with the full collected set of articles yet and during the last year of the project we expect to improve the platform and fulfil this lack of performance by being able to analyse all the collected articles in real time.

During the first year we also sketched a plan for reaching targeted audience through different channels in order to gain awareness in the context of language technologies, text mining, and cross-lingual information extraction. An updated plan including the showcase scenario and the new functionality implemented during the second year towards filling the gap between industrial and academy fields has also been drawn and it is presented at Section 4.

Finally the conclusions are presented at Section 5 including the future steps to deal with the major pending issues that are expected to be covered during the third year of the project.

2 XLike Showcase Scenario

This section includes the detailed news agency story related with the industrial showcase which was firstly analytically detailed in [5] as shown in Table 1. This story introduces the different steps that an editor of a press agency has to go through in order to be able of being the first in finding an event that is currently happening or has just happened worldwide. It is worth to pinpoint that this industrial use case has been validated with the STA Slovenian Press Agency [4], but it can be seen as representative and applicable to any other press agency since they all operate in this part of news production process in similarly. At different steps of the story the tasks that are being used according to Table 2 are also indicated.

Table 1. Summary of the industrial showcase specifications.

	Cross-lingual articles/news tracking for event discovery
Application	Cross-lingual event detection and linking
Input	a) Mainstream news stream b) Social media stream
Output	Web-based graphical tool for monitoring current events and their corresponding articles from all XLike languages.
Tasks providing tools for the showcase demonstrator	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news and social media input for event detection and article tracking. **T2.1 – Shallow linguistic processing** of standard language used for language identification, tokenization, lemmatization and named entity extraction. **T2.2 – Deep linguistic processing of formal language** deep processing required for relation extraction and creation of semantic graphs. **T2.4 – Extracting structure from informal language corpora**, extending the coverage of showcase demonstrator to less standard language, such as in Twitter tweets. **T3.1 – Approximate text annotation with cross-lingual semantic repositories**, providing semantic context to extracted entities and relations. **T4.1 – Statistical cross-lingual document linking** used to identify related articles across language barrier. **T4.2 – Semantic graph construction** will showcase how output from WP2 and WP3 can be combined to create rich semantic graph representations of news documents. **T4.3 – Event extraction for semantic graphs** used for the definition of event templates needed for the event detection, and their population from multilingual news feed. **T5.2 – Information visualization** will be used as primary GUI components for the showcase, and will be used to show the detected events and the associated news in different languages. **T6.2 – Integration platform** which will host all the functionality needed providing enough flexibility and scalability. **T6.3 – API for exploratory real-time data stream analysis** to provide a scalable exploratory analysis over large social and new media data streams. **T6.4 – Desktop and Web front-end** to provide a quick prototyping user interface to provide access to the existing functionalities.

2.1 The Early Event Detection Story: Bob the editor

1. (September the 3th, 2013 early morning) Bob, a business news desk editor, is at home and he finds out via Twitter unconfirmed blogger information saying that the Telefonica, a Spanish telecom company, will face a stronger competition in the German market, as British operator Vodafone is taking over the German biggest cable operator Kabel Deutschland. The blogger only mentions that the information comes from an unspecified Spanish media.

2. Meanwhile the XLike platform has been monitoring in real time 24x7 many different sources and Twitter is one of them. The platform provides multi-lingual and cross-lingual quick access

functionalities through a user friendly interface which allows access and search capabilities over large volume of data and establishes links with similar news appearing in different languages. Therefore, this unconfirmed and unofficial information about Telefonica should have been tracked by the platform and for sure the platform can provide some insights over its origin and trustworthiness. This step is related with T1.3, T2.1, T2.2, T2.4, T3.1, T4.2, and T6.2.

Searching for the story

3. When Bob gets into the office he opens the XLike web application and selects the event registry tool searching for the information that he read before about Telefonica. He uses the event registry tool[1] and searches for Telefonica and Vodafone events which have been published during the last week in Spain. Since there are many stories related with these two companies, he has to narrow down the search to business category using the XLike user friendly interface. This step is related with T4.3, T5.2, T6.3, and T6.4.

4. Now Bob can observe among the top results the event that he is looking for. Then, he opens the event and chooses only Spanish articles. He is interested in knowing which media house was the first to publish the information in order to verify its authenticity. By clicking on the sorting button all the selected articles are sorted by their publication time. This step is related with T5.2.

5. When Bob finds out that the first media house which published the information did not contain any official confirmation, he goes through all the other articles trying to figure out if that information can be trusted or not. Then, he finds out an article also from a more prominent Spanish news source which contains some official information. This step is related with T5.2.

6. Because Spanish media predominantly only include Spanish official sources, the editor wants to find out also articles from another countries as Germany and Great Britain. For this purpose he goes also through the tabs with articles in English and German language to obtain related articles of the same story. This step is related with T2.1, T2.2, T3.1, and T6.4.

Editing the story

7. After this first search Bob decides to write a story for his media house. As he wants to base the story to his most trusted sources, he makes the search for the event also by publisher. This step is related with T6.4.

8. When he finds the articles from the trusted sources, he starts to write the story. Since the real time coverage is important in publishing news outbreak, Bob has to be as fast as possible and also has to be sure that he mentions all the important aspects of the event. For that purpose he opens the story and takes a glance at the top named entities and keyword lists and reuses them for the writing. This step is related with T4.3 and T5.2.

9. Since the events are never isolated, Bob wants to give also some more background information and history to his story. He uses the tabs "similar events" and "to find out some additional information" and also links to discover related news. This task is related with T4.1.

Publishing the story

[1] http://eventregistry.org/

10. After publishing his story, Bob uses the Article browser tool to find out how fast he was in publishing his story compared to other Slovenian media houses. This task is related with T1.3, T6.2, and T6.4.

Bob was happy because he has been the first one by using the XLike platform!

This industrial showcase story collects many of the desired capabilities of any industrial showcase within the news streams and social media context and also uses the developed capabilities of the XLike project including the cross-lingual ones. Other industrial stories where these functionalities can be applied are *brand reputation* by automatically gaining information about companies through searching for events related with them, or *web mashups* by adding information of the events related with specific categories, concepts, or keywords, (e.g. answering to the question, what is being said around my company?). Tools like NetBase[2] or SocialMention[3] provides insights into sentiment and trends about a brand, or mashups of information but they refer and manage the data as isolated pieces of information.

The above presented story belongs to a new industrial trend in data journalism for helping journalism sector in an early detection of events (e.g. many times journalist speaking a lesser spread language have to go to other countries/languages to obtain the initial information. Also applies in the opposite case when a journalist seeks an event that happened in place where lesser spread language is used and s/he needs to consult the sources at that language.). The introduced story also makes use of the different components of the project as it is indicated at each step and they have been designed following a SOA approach and independently per language which allows scaling it as needed (e.g. due to new source media being collected, or larger number of published articles, etc.) by adding additional computers into grid.

The SOA approach provides also the opportunity to offer the functionalities provided by the project as SaaS (Software as a Service) fulfilling both: i) make the XLike toolkit to survive beyond the end of the project and also ii) to easily scale the functionalities (including the cross-lingual event detection which is the one presented here) to achieve real time performance.

Some general requirements which have been demonstrated in this industrial showcase are:

i) creation of a Toolkit which is easily deployable and can be scaled by size and language allowing the transfer and creation of new platforms replicating the original,

ii) execution of the different language pipelines almost in real time being able to analyse one third of the collected data,

iii) analysis of informal social media, where often a non-standard variety of language is used, such as Twitter. This analysis performance is running in near real time, and

iv) visual means for easy interpretability of the results.

In the next section we describe the software and architectural features that have been implemented during the year two of the project and that are available in the XLike platform so far.

[2] http://www.netbase.com/solutions/social-brand-analysis/
[3] http://socialmention.com

3 Components

This section covers the specific components used to accomplish the above presented showcase scenario. The XLike components, which cover the needed functionality in order to provide a final fully functional demo, draw heavily on the Final *toolkit architecture* specification [1] and Requirements for demonstrator [2].

3.1 Showcase software/architectural features

In order to accomplish with this showcase some technological and execution decisions were taken. These decisions have been also taken considering the showcase software/architectural requirements defined in [3].

The technological requirements which have been accomplished are:

- The used platform contains different REST services allocated at different places in a SaaS approach. This makes the platform extremely flexible allowing fast incorporation of new functionalities and also allows the control of load balance. The fact that services are not allocated physically at the same place is not a major drawback due to the messages between the different services of the platform contain only texts which are not very large in volume.
- The different pipelines have been implemented independently of each other for allowing its parallelization. So far each pipeline supports multiple forks/threads and they are also deployed at different number of computers allowing their easy scalability.
- The filtered data provided from different resources and the data generated as result of the showcase have been stored and are accessible via web services[4] and via web site[5].
- The multilingual pipelines produced within XLike and functioning as web services have been shared with the LT community through the META-SHARE platform.[6] In this way their accessibility for the research and other communities is preserved beyond the end of the project.
- The interaction with users has provided by a simple and fully functional web site[7] that displays the report of the detected events in the associated news/articles written in different languages.
- The interface has been implemented using JavaScript and JQuery, and can be easily deployed as another component of the platform. It also provides direct and integrated linked information to external sources such as Wikipedia or DBpedia for easier interaction and guides the searching.
- Map, timeline, and hot trending metaphors and specific visualization templates have been implemented for easier interpretability and gaining bird-eye insights into the stories and events detected.

The different and independently deployed components are constantly being tracked by monitoring tools[8,9] in order to avoid cascading of failures and make the platform more robust. This monitoring tool provides a quick alert whenever a service is down in order to check them out. Furthermore some internal controls have been implemented for each service for allowing automatic re-starting of services. The components also are independent of each other allowing their exclusion avoiding the cascading of failures making the platform more robust.

[4] http://eventregistry.org/?query=
[5] http://eventregistry.org/
[6] http://meta-share.eu
[7] http://eventregistry.org/
[8] http://stats.pingdom.com/33sj6sc6keuh
[9] https://www.leftronic.com/share/9tcE6X/#9tcE6X

There is still one requirement which was not accomplished so far which is related with making the current platform easily extensible to a higher number of computers or to distributed systems (e.g. Amazon Elastic Compute EC2[10]). In this direction we have started to work towards having an easy deployable platform by migrating some of the current components of the JSI NewsFeed to open source solutions (e.g. Apache Cassandra[11]) which are better known by the technological community and also allows to plug-in other tools which provide extra distributed capabilities (e.g. Apache Hadoop[12]).

Currently, the showcase uses QMiner platform [7] for its backend, which is an analytics platform for large-scale data stores and real-time streams containing structured and unstructured data. It is designed to for scaling to millions of instances on high-end commodity hardware, providing efficient storage, retrieval and analytics mechanisms with real-time response. At the moment QMiner is using a data storage functionality developed at JSI, which is designed to work on a single node. As we mentioned above, during XLike project we plan to implement support for Apache Cassandra, which would allow us to scale the data layer to a cluster of machines to facilitate scaling. We decided for Apache Cassandra due to high write throughput.

3.2 Showcase – XLike Event Identification/Registry

In this section we describe the overview of the different components used for the above described event registry showcase and we provide details of the different functionalities that it offers to the end-users. For that purpose we introduce a summary of the different components and the architecture of the showcase, and how they accomplished with the expected outcomes of the [5].

Table 2 Industrial showcase specification functionality at D8.2.1

	Cross-lingual articles/news tracking for event discovery
Application	Cross-lingual event detection and linking
Input	a) Mainstream news stream b) Social media stream
Output	Web-based graphical tool for monitoring current events and their corresponding articles from all XLike languages.
Tasks providing tools for the showcase demonstrator	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news and social media input for event detection and article tracking **T2.1 – Shallow linguistic processing** of standard language used for language identification, tokenization, lemmatization and named entity extraction **T2.2 – Deep linguistic processing of formal language** deep processing required for relation extraction and creation of semantic graphs **T2.4 – Extracting structure from informal language corpora**, extending the coverage of showcase demonstrator to less standard language sources, such as in Twitter tweets **T3.1 – Approximate text annotation with cross-lingual semantic repositories**, providing semantic context to extracted entities and relations **T4.1 – Statistical cross-lingual document linking** used to identify related articles across language barrier **T4.2 – Semantic graph construction** is being used to determine and extract semantic facts from a group of articles on the same event **T4.3 – Event extraction for semantic graphs** used for the definition of event templates needed for the event detection, and their population from

[10] http://aws.amazon.com/ec2/
[11] http://cassandra.apache.org/
[12] http://hadoop.apache.org/

	multilingual news feed **T5.2 – Information visualization** will be used as primary GUI components for the showcase, and will be used for visualization of the detected events and the associated news in different languages **T6.2 – Integration platform** which will host all the functionality needed providing enough flexibility and scalability **T6.3 – API for exploratory real-time data stream analysis** to provide a scalable exploratory analysis over large social and new media data streams **T6.4 – Desktop and Web front-end** to provide a quick prototyping user interface to provide access to the existing functionalities

3.2.1 Event registry architecture

The event registry prototype is based on the large set of tools/components that have/are being developed within the XLike project and which have been documented in D6.1.2 “Final Toolkit Architecture”. The basic architecture is shown in Figure 1 and a detailed description of all the different components and modules that it contains are explained in detail at D4.3.1 “Early event extraction prototype”. In this document we provide an end user and showcase perspective of that functionality and how it was accomplished with the initial specifications.

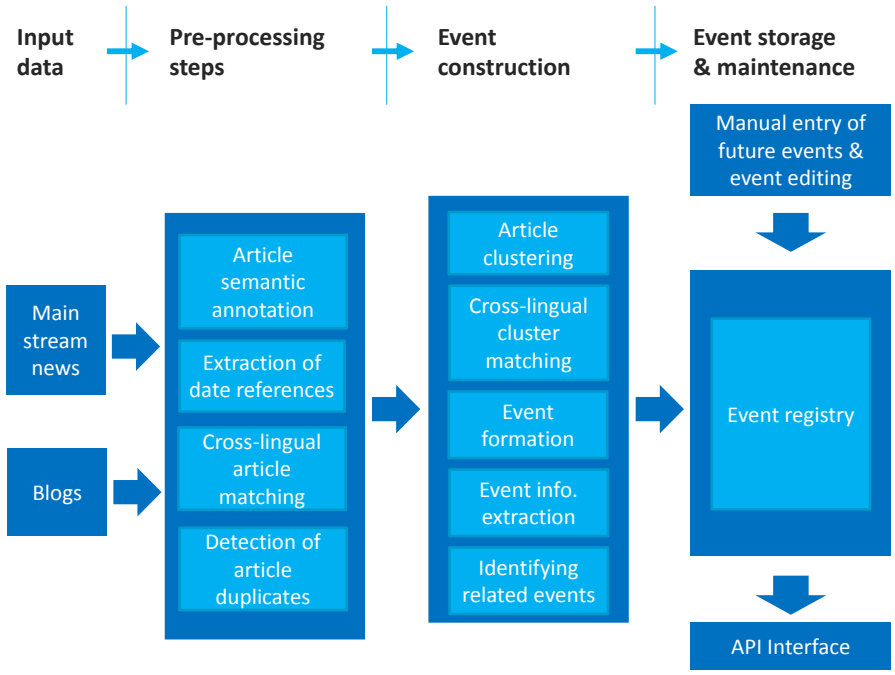


Figure 1 Pipeline used for event detection and their storage

3.2.1.1 Raw input data

The data is being collected by JSI NewsFeed (accomplished with T1.3 at Table 2) and it consists of main news media sources and popular blog sites. The total number of different RSS feeds is around 75.000 and generates between 100,000 and 150,000 articles per day. Most frequently represented languages are English (50% of articles), German (9%), Spanish (8%), French (5%), and Chinese (4%). This data is being used to find relevant events after the processing.

3.2.1.2 Pre-processing the data

In order to analyse the collected articles, each one is passed through a series of pre-processing steps. These steps include shallow and deep linguistic processing (accomplished with T2.1 and T2.2 at Table 2),

annotation of the text with cross-lingual semantic resources (accomplished with T3.1 at Table 2) and cross lingual document linking (accomplished with T4.1 at Table 2). The annotations provide a list of entities and keywords which are mentioned in the text, and the cross lingual document linking is performed using canonical correlation analysis (CCA) [6] which provides a list of similar articles in other languages and an estimation of their similarity. For the purpose of event information extraction date references which are mentioned in the articles are also identified.

3.2.1.3 Event construction

An event is defined as a set of articles which talk about the same fact. For finding those sets of articles a clustering algorithm for identifying groups of articles that describe the same event is applied after the pre-processing step. The clustering algorithm uses as input features article text, the identified concepts (entities, keywords, semantic roles) and article date, when deciding which articles to group. Since events are often written about in different languages, there is also a need of linking groups of articles on the same event in different languages. For that purpose, the information obtained from cross-lingual article linking (T4.1), the article annotations (T3.4, which are language-independent), date similarity, and other features extracted automatically from each group of articles are used. After one or more groups of articles are identified, it is believed that they are representing the same event and a new event is added to the event registry (accomplished with T4.3 at Table 2).

Then for each group of articles associated with the event an inverse engineering process is performed in order to extract relevant information associated to that event. This information allows categorizing the events by using frequently occurring annotations which highlights the relevant concepts and topics. Moreover, if a location is identified frequently enough (especially at the beginning of the article) then it is considered as the location of the event. Similar procedure is performed with the date and if no such date is found, then the average publishing date of the article is used as the event date.

In order to provide richer information and enhance the searching capabilities we have considered that the events can be about different topics, such as meetings, earthquakes, or bombardments. In order to classify the topic of the event the DMoz[13] taxonomy has been used. Then, for each event we combined the text from the articles and apply on it a DMoz classifier assigning it to the category with the highest probability.

3.2.1.4 Event Storage

The detected events are stored in the event registry database-like storage (accomplished with T6.2 from Table 1). The events are indexed across different fields which allow the users to search for them using different criteria via publicly accessible APIs. Also, the frontend was developed to provide search functionality as well as various ways to visualize and aggregate the search results. The core functionality of this frontend is part of the event registry user interface and it is detailed in the next section.

3.2.2 Event registry user interface

The end-user interface is split into two main parts which allows the access to all functionalities implemented: i) search interface, and ii) event information interface (this is accomplished with T6.3 and T6.4 from Table 1).

3.2.2.1 Search options

This visualization helps users to find the events they are interested in. The event registry has to offer an extensive set of search options for empowering its usability. The currently supported search options are shown in Figure 2. It provides searching by concept or by location which can be either a city or a country. Other search parameters that can be set are the publisher, time of interest, limit of the results obtained,

[13] http://www.dmoz.org

and the chosen categories according to DMoz. The search can also be constrained by the minimal number of articles covering the event.

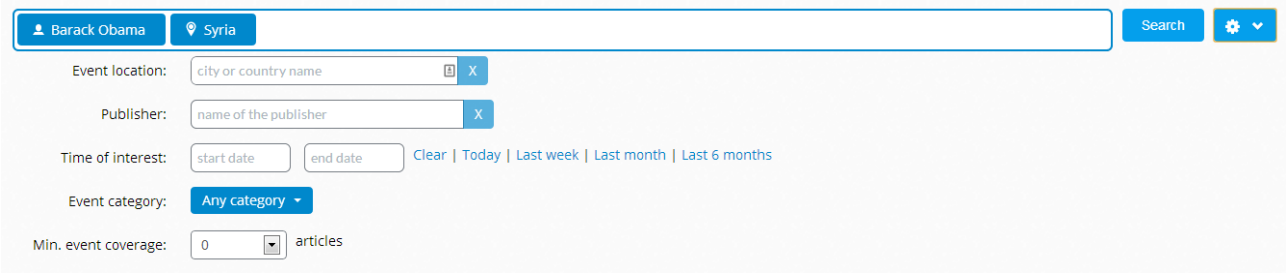


Figure 2. Event registry search options

3.2.2.2 Displaying search results

The visualization of the search results provides an easy way to better understanding and possibly refinements of the query. The most common way to present the results is as a list of events (see Figure 3). For each of the events the main extracted event information, including the title of the event, date, location, main entities and keywords, number of articles reporting the event, and a short summary is shown. The events can be ordered either by relevance to the query (most relevant first) or by date. By clicking on the title of the event a new window is opened displaying the details of the event (see displaying event information in the next section).

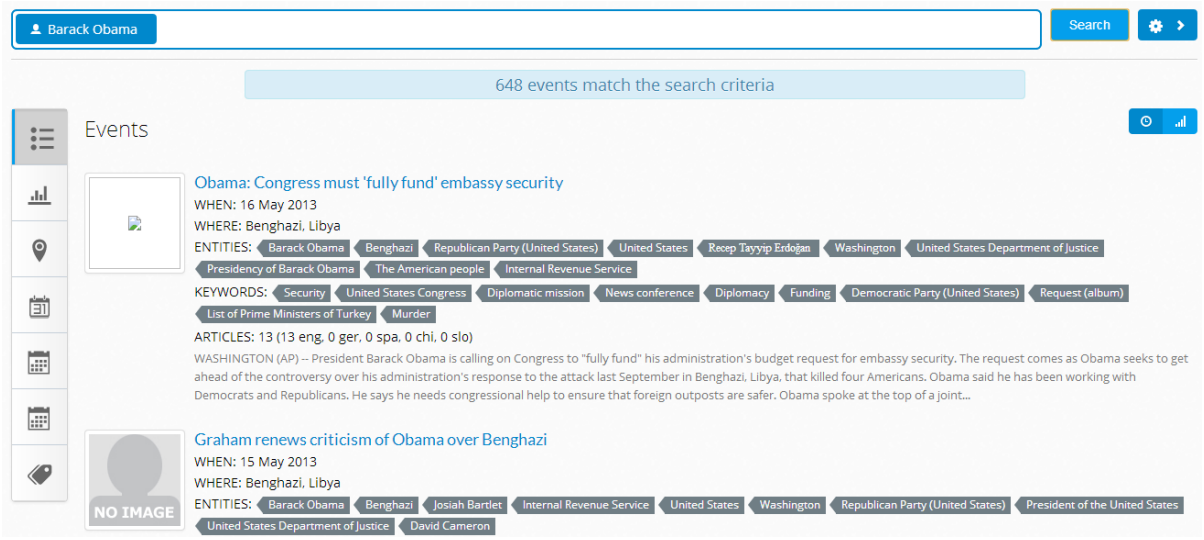


Figure 3. List of events

Furthermore, in obtaining the big picture of the search results (including possible analytics), a set of visualizations have been implemented in form of bar charts displaying the most relevant concepts (see Figure 4). Similar visualization is performed for the location and time (see Figure 5, right and left) (this is accomplished with T5.2 and T6.4 at Table 1).

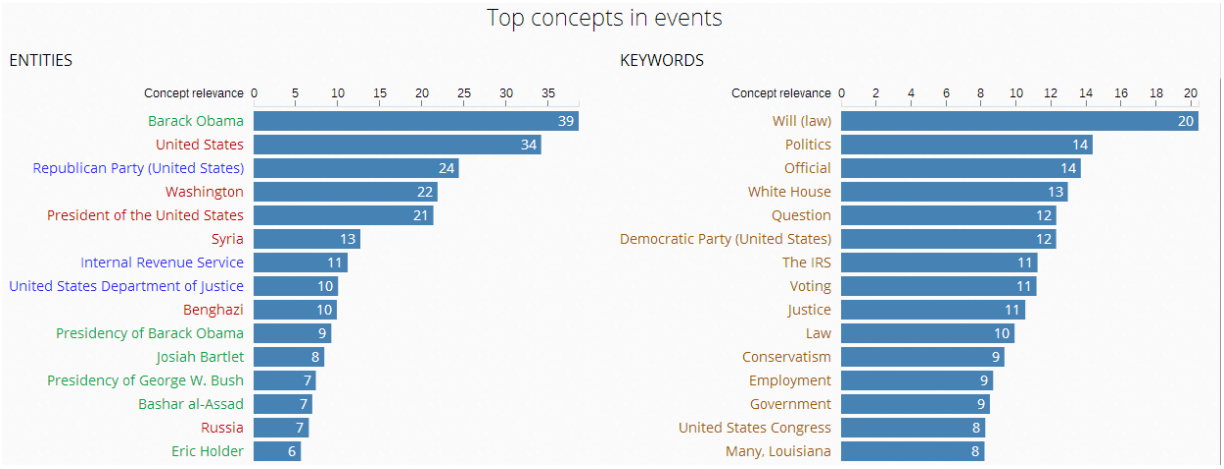


Figure 4. Top concepts and entities that occur in the results

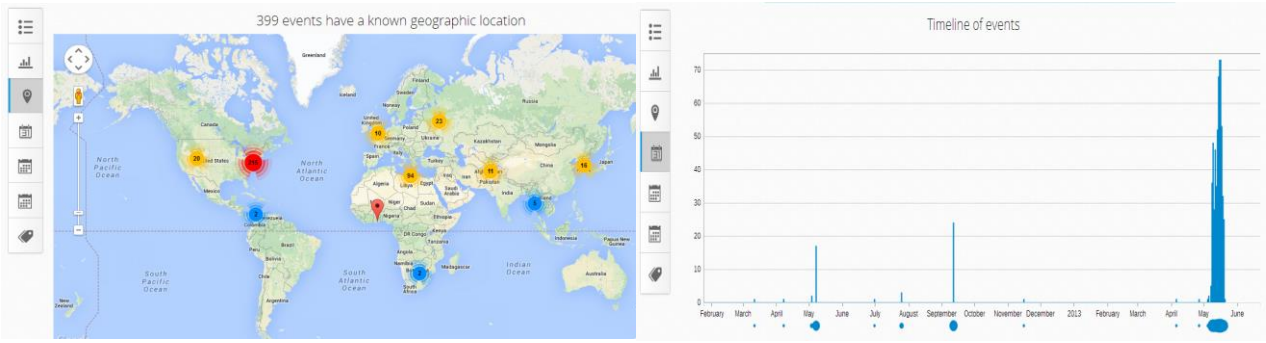


Figure 5. Geographic location and time distribution of the events

Also, the trending concepts graph (Figure 6) can be used to see how the popularity of top concepts changes over time providing a guess about its interest peak.

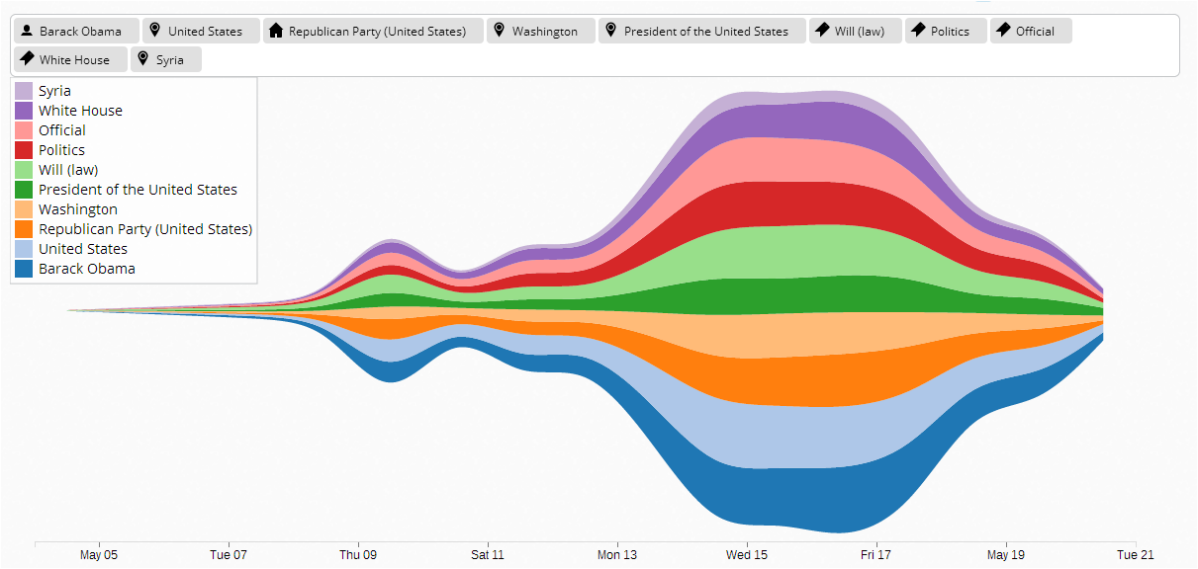


Figure 6. Trending of top concepts over time

3.2.2.3 Displaying event information

The editor needs to write the story once he has detected something of an interest and has searched for it using the search capabilities. Figure 7 shows the event information. At the top it presents the title, location, date and a short summary of the event. Below, the list of articles dealing with that event is placed. The articles are organized by language and this can be changed by selecting the desired language. By clicking the title of the article, the actual content of the article can be consulted.

There are two other relevant views related to the event. The first is the timeline of articles reporting about the event. Figure 8 shows the graph that allows discovering when the event of interest was first reported and how were the reports trending over time. The height of the curve indicates the cumulative number of articles about the event in the last 6 hours.

This and the previous time and location distribution are useful within the presented story in order to check if the editor has been on the initial rising part of the peak or in the decay (between the first mentioning of it or between the last ones).

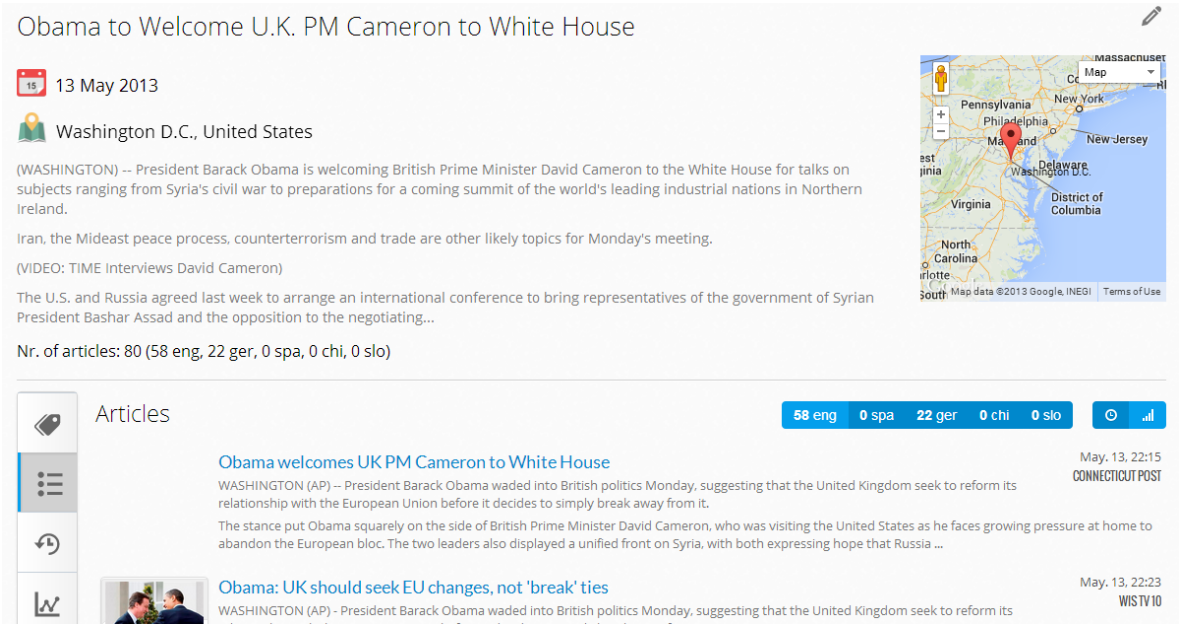


Figure 7. Event details and the list of articles reporting about it

Besides, while tracking the coverage of the event, we can also identify other related events (see Figure 9). This can enable the editor to broaden the perspective on the event itself. Similar events are identified by usage of the event's concepts at similar relevance level, i.e. by finding other events with a similar "signature".

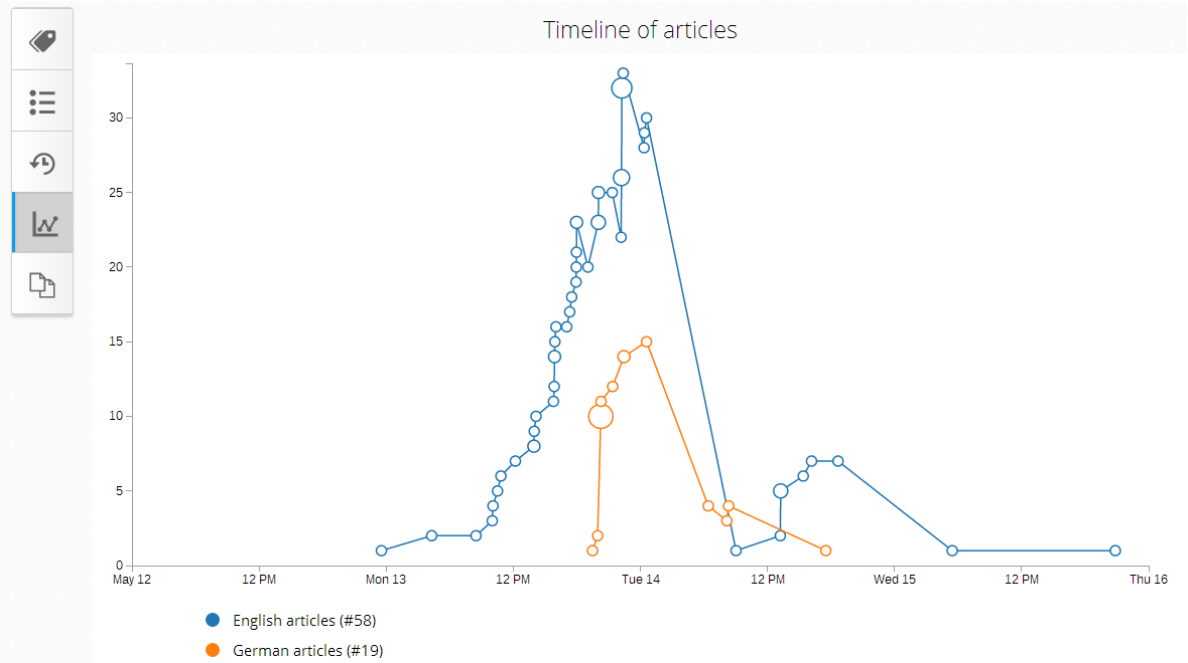


Figure 8. Trending of articles in different languages reporting about the same event

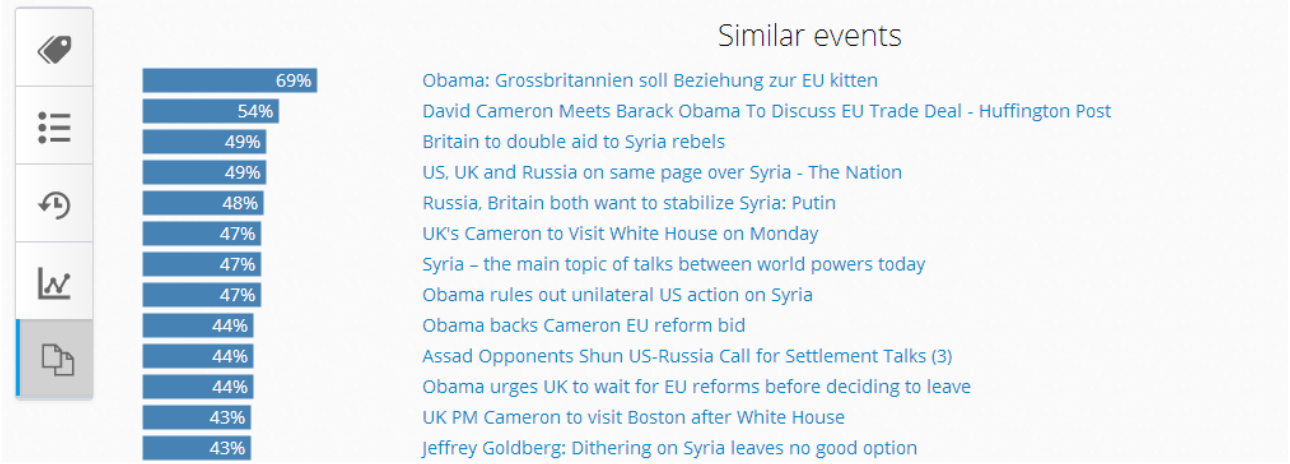


Figure 9. The list of other similar events including their similarity based on concept agreement

4 Presentation requirements and materials

In the D8.1.3 Communication plan, we have defined our general dissemination strategy. What is relevant for this deliverable, apart from the general dissemination actions (visual identity; usage of predefined templates for PPT, posters etc.; usage of XLike web site for promotion; etc.) is the definition of the audience target groups XLike project is aiming at.

XLike dissemination activities are focused on the following **target groups**:

- **Scientific and research community**;
- **Industry and customers**;
- **General public**.

In this respect the target group we are interested in here, is the **Industry and customers** – companies and professionals as potential users of XLike technologies interested in improving the quality of their news production process and interested in applying sophisticated language technology and semantic processing in their production process. For them we have introduced this industry outreach showcase that will be used to demonstrate XLike technology platform at different conferences and industry gatherings.

In addition, education/training in the form of workshops and summer schools for members of research and industry communities was also one of our dissemination channels in the second year of the project.

4.1 General presentation material: XLike flyers

The initial XLike flyer has been produced for dissemination of the project aims and goals, primarily at the important scientific events in 2012 (e.g. LREC2012, EAMT2012, KTE2012, etc.). In the meantime, a mid-term flyer has been produced and it will be used also in the industry showcases as a general introduction to the XLike project while presenting the achieved progress so far.

XLIKE

Cross-lingual Knowledge Extraction

Project partners

IJS: Jožef Stefan Institute, Ljubljana, Slovenia
KIT: Karlsruhe Institute of Technology, Karlsruhe, Germany
UPC: Universitat politecnica de Catalunya, Barcelona, Spain
UZG: University of Zagreb, Zagreb, Croatia
TSINGHUA: Tsinghua University, Beijing, China
ISOCO: Intelligent Software Components S.A., Madrid, Spain
BLOO: Bloomberg, New York, USA
STA: Slovenian Press Agency, Ljubljana, Slovenia

Associated partners

NYT: New York Times, New York, USA
IIT: Indian Institute of Technology, Bombay, India

Contact

Artificial Intelegence Laboratory
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
SLOVENIA
Contact person
Mojca Kregar Zavrl, project manager
P: + 386 1 477 3853
E: mojca.kregar@ijs.si
W: http://www.ailab.ijs.si

www.xlike.org

XLIKE

Cross-lingual Knowledge Extraction

SEVENTH FRAMEWORK PROGRAMME

Funding: The project has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)
Area: Language Technologies (ICT-2011.4.2)
Project reference: 288342
Total cost: 4.57 Meuro
EU contribution: 3.55 Meuro
Duration: from January 2012 to December 2014 (36 months)
Contract type: Small and medium scale focused research project (STREP)
Coordinator: Marko Grobelnik, Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana, Slovenia

XLike Leaflet 2013-07

www.xlike.org

XLIKE

Cross-lingual Knowledge Extraction

XLIKE

Cross-lingual Knowledge Extraction

Main goal

The goal of the XLike project is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence.

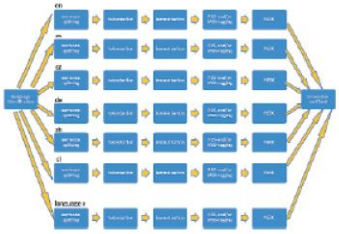
Research contributions

The effort combines scientific capabilities and insights from several areas of science – computational linguistics, machine learning, text mining and semantic technologies – in order to enable cross-lingual text understanding by machines.

Specifically, we plan to pursue the following two key open research problems:

- to extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases;
- to adapt linguistic techniques and crowdsourcing to deal with irregularities in the informal language used primarily in social media.

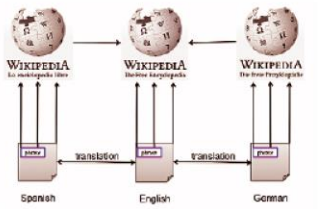
The developed technology is language-agnostic, while within the project we specifically address



English, German, Spanish, Chinese and **Hindi** as major world languages and **Catalan, Slovenian** and **Croatian** as minority languages.

Knowledge representation

Knowledge resources from **Linked Open Data** cloud will be used, with special focus paid to using general common sense knowledge base **CycKB** for Interlingua. For languages where no required linguistic resources are available, we will use a probabilistic Interlingua representation trained from a comparable corpus derived from the Wikipedia and advanced methods in Machine Translation.



XLIKE

Cross-lingual Knowledge Extraction

Results achieved so far

In the first year of the project several goals have been achieved:

- linguistic (pre-)processing pipelines for six languages using the same methodology;
- cross-lingual document linking using Canonical Correlation Analysis (CCA) alowing training of similarity models for millions of documents from Wikipedia;
- automatic cross-lingual annotaton and linking of NEs and concepts from texts to Wikipedia articles.

Use cases

The early prototypes integrating all the technology developed during the first year were deployed and tested with the industrial patners:

- **Bloomberg** use case, covering the domain of financial news where a specific measurement of news relevance for Bloomberg was used;
- **Slovenian Press Agency** use case, covering the domain of general news.

Figure 10: The XLike mid-term flyer

For industrial showcase a limited number of special purpose flyer will be produced where that showcase will be presented as demonstration at different conferences during 2014. XLike industrial partners will be involved in the production of these promotional materials, particularly its content.

4.2 Industry outreach

We find very important that XLike project and its technological results are presented to industry and that industry is aware of the project and its achievements. The preferred aim of these activities is that industry adopts the outcomes of the project and to turn them into products and services. The use cases and the proof-of-concept prototypes are valuable instruments to introduce XLike at industry conferences and business and technology gatherings.

Actions and instruments undertaken for the showcase demonstration in the second year of the project are:

- mid-term flyer: providing the general view of the project and its achievements until mid 2013.
- dedicated showcase flyer: providing a showcase scenario (editor Bob) describing the specific approaches and technological solutions in his news production process
    - linguistically and semantically motivated approaches (pipelines that feature shallow and deep linguistic processing, semantic role labelling, mapping to ontology relations, usage of Cyc/LOD, machine translation as aiding technology...)
    - statistically motivated approaches (cross-lingual detection of similar documents, CCA and ESA approaches,...)
- specific points of interest for the industry have been defined and will be used for demonstrations in the form of prepared FAQs
    - Harvesting: How we harvest data? How we make it available for industrial purposes? How we solve IPR issues of web collected data?
    - Processing speed: How we optimise the processing in XLike Toolkit?
    - Processing efficiency: How we scale up the background computing facilities, i.e. grid? What are typical computing scenarios and which computing resources have to be allocated for their processing?)
    - Processing tuning: How we tune up the systems for precision and/or recall? How to balance the system towards this measure to meet the industrial requirements and to achieve the best performance? How to tune the system for specific user needs?

These activities will be undertaken mostly after M24 since for this kind of target audience we will have to have a working prototype that clearly demonstrate the benefit of usage of the technology that XLike is developing. In this respect the dedicated showcase flyer is targeted for M26 when we expect the series of demonstrations will begin.

Already we have created several videos which show the main functionalities related with this industrial showcase and also the following material has been released and tested:

- event registry introduction, event registry search, and event registry information[14].
- near-industry strength processing pipelines (see D6.2.2 “Demonstrator prototype” and multilingual demo[15]).

[14] http://eventregistry.org/intro
[15] http://sandbox-xlike.isoco.com/demo/index

- the pipelines are independent and can be used separately for different purposes (for instance French and Italian have been added easily to the pipeline for the Bloomberg use case which shows the adaptability and independency of the whole process and technology).
- the near industry processing pipeline functionalities have been used sucessfully to the large streamlines of XLike industrial partners (BLO, NYT, STA...) in XLike languages
- the near industry processing pipeline functionalities have been also used sucessfully to social media streamlines (Twitter, Facebook, blogs...) in XLike languages

In this respect we are aiming to attend and actively promote XLike at two types of events.

4.2.1 Industry conferences and meetings

During the first year we provided a set of conferences which would be of interest in order to aware industry partners and find out new industry showcases related with XLike. Among others we proposed ESTC, SemTech, TextAnalytics, I-SEMANTICS, and industrial Workshops organized by major international conferences (such as ICDM or KDD).

In this respect XLike has already presented this industrial showcase to industrial companies such as Xinghua news agency (at Tsinghua University on September the 11th), has given a class conference "From News to Events" (invited talk at Bled on October the 22nd), and a seminar "Detecting events from news articles (AI department at JSI on November the 6th).

Although due to the industrial showcase has been recently completed we have already submitted it to the demo track at WWW'2014[16]. Furthermore we have also submitted to the same tack the overall demonstrator of the XLike project as results of the work done during the second year which includes the multilingual processing pipelines as a news agency tool showcase.

Two XLike demonstration papers have been sent to LREC2014 conference (deadline in 2013-10), one is planned for EACL2014 (deadline 2014-01), one for ACL2014 (deadline 2014-02), one for COLING (deadline 2014-03). The XLike project will be one of official organizers of EAMT2014 (2014-06).

4.2.2 Awareness and networking events

Despite some actions have already been taken in this direction, most of the awareness and dissemination will take place during the third year of the project. For this purpose XLike has prepared the following demos and showcases:

- Multilingual pipeline demonstrator[17].
- XLike Demonstrator (includes the multi/cross lingual functionalities, visualization, and searching capabilities focused on providing a tool for easy access to information at different languages)[18] .
- Event Registry/search tool[19].
- Cross-lingual Articles Recommendation[20].

So far XLike has organized a presentation of the project results on several occasions, e.g. at META-FORUM (Berlin, 2013-09), ELRA workshop *Sharing Language Resources: Landscape and Trends* (Paris, 2013-11) etc., but new events related with developing functionalities using the currently implemented APIs and also for

[16] http://www2014.kr/
[17] http://sandbox-xlike.isoco.com/demo/index
[18] http://sandbox-xlike.isoco.com/portal/
[19] http://eventregistry.org/
[20] http://aidemo.ijs.si/xling/wikipedia.html

showing the demos and showcases will be held mostly during Y3 to close the loop towards near-to-market solutions.

4.3 Training for professionals

The focus on this task is to target companies which can potentially be early adopters of the XLike technologies. For that purpose the work done so far has been focus is providing a set of demos and tutorials which can be used for highlighting the capabilities and advantages of a cross-lingual knowledge extraction empowered by the usage of XLike sophisticated mining methods, lightweight semantics, and powerful natural language processing.

During the Y2 XLike partners have done the following actions related with the industrial awareness and training:

- Presentation of the Y1 early prototype at CeBIT[21] which is the world's leading event for digital business.
- XLike project coorganized the Summer School in Dubrovnik (2013-10-07 – 2013-10-10) where the project was presented to early-stage researchers oriented towards research in industry.
- Online demos and videos for the core functionalities of the project.
- Tutorial examples[22] for developing and accessing the XLike public API's and XLike Toolkit.
- Presentation at the European Business Press Publishers' seminar in Stockholm (2013-11-14).

During the third year we expect to continue extending the tutorials and videos (making them public through videolectures.net at JSI), and also organizing some hands-on seminars to promote the use of XLike technologies.

[21] http://www.cebit.de/home
[22] https://github.com/xlike-project/XLikeTutorial

5 Conclusions

This document has presented the implementation of the industrial showcase as result of the work done during the second year of the project towards fulfilling the requirements and functionalities described in the showcase specification scenario [5] during the first year (M6). This showcase scenario has been grounded by defining a specific story which assists journalists for early detection of events in order to help them through the process of writing an article in a news social media environment.

The journalist assistance for early detection of events is one of the most important steps of news' production processes due to the high competition for being the first publishing a piece of news. XLike project offers help in this task by providing a set of innovative tools and functionalities (WP1-WP5) which have been grouped to build up a toolkit which offers multi and cross lingual functionalities for several languages (English, German, Spanish, Catalan, Slovenian, and Chinese) and have been used to accomplish with the "early event detection story".

This deliverable, the implemented components, and the event registry tool are the result of the work done towards providing a closer real life scenario where the generated technology can be deployed for helping actual end-users in news agencies (in our case STA). We also have envisioned some other domains such as brand reputation or mashups where the use of events as basic information units instead of independent treatment of sources of information could be relevant and may provide some industrial advantages such as broader coverage and contextual information .

A set of promotional material for industrial awareness and outreach has been reported and also new initiatives have been defined for the third year where this showcase will have major impact due to it will be fully available. Furthermore special emphasis will be put into filling the gap between the industry and the academia by presenting these results at industrial workshops, demo tracks, and through tutorials.

References

[1] D1.2.2 – “Requirements for demonstrator”

[2] D4.3.1 – “Early event extraction prototype”

[3] D6.1.2 – “Final toolkit architecture specification”

[4] D7.2.2 – “Demonstrator and validation report”

[5] D8.2.1 – “Showcase specification”

[6] M. Borga, O. Friman, P. Lundberg, H. Knutsson; “A Canonical Correlation Approach to Exploratory Data Analysis in fMRI”, ISMRM 2002, Honolulu, Hawaii, USA, May 2002

[7] D6.3.1 – “API specification and prototype”

