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

Deliverable D8.2.1

Showcase specification

Editor:	Esteban Garcia-Cuesta, iSOCO
Author(s):	Esteban Garcia-Cuesta, iSOCO; Marko Tadić, UZG; Blaž Fortuna, JSI
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)[1]	Public (PU)
Contractual Delivery Date:	M6
Actual Delivery Date:	M6.5
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

[1] Please indicate the dissemination level using one of the following codes:
• **PU =** Public • **PP =** Restricted to other programme participants (including the Commission Services) • **RE =** Restricted to a group specified by the consortium (including the Commission Services) • **CO =** Confidential, only for members of the consortium (including the Commission Services) • **Restreint UE =** Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments • **Confidentiel UE =** Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments • **Secret UE =** Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

Disclaimer

This document contains material, which is the copyright of certain XLike consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):
All XLike consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):
All XLike consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):
The information contained in this document is the proprietary confidential information of the XLike consortium and may not be disclosed except in accordance with the consortium agreement. However, all XLike consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):
The information contained in this document is the proprietary confidential information of the XLike consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the XLike consortium as a whole, nor a certain party of the XLike consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

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:	D8.2.1 – XLike showcase specification
Editor (Name, Affiliation)	Esteban Garcia-Cuesta iSOCO
Work package Leader (Name, affiliation)	Marko Tadić, UZG
Estimation of PM spent on the deliverable:	

Copyright notice

© 2012-2014 Participants in project XLike

Executive Summary

This document presents the first specification of the industrial showcase which will be accomplished during the development of the XLike project and will be demonstrable by a functional demo. This work also includes promotional material to be used for dissemination purposes and diffusion of results.

This deliverable is the first of two where the initial specifications for the showcase are provided whereas in the second (D8.2.3 XLike Showcase) the final fully functional demo will be ready for dissemination, commercial awareness and use of the European Commission.

This document contains the definition of the scenario to be used as a working proof for the M24 and the software requirements which will be needed for the successful development of a demo applied to that scenario. The set of components that the project will provide are also specified including its precedence and its place in the overall pipeline of the XLike project.

The document also contains the description of some of the promotional material that has been created for the dissemination of the XLike project and more specifically for industrial awareness.

© XLike consortium 2012 - 2014

Table of Contents

Executive Summary 3
Table of Contents 4
List of Figures 5
List of Tables 6
1 Introduction 7
2 Industrial showcase 8
3 Software Requirements and Components 9
3.1 Showcase software/architectural requirements 9
3.2 Showcase – XLike required components 9
4 Presentation requirements and materials 13
4.1 General presentation material: XLike flyers 13
4.2 Industry outreach 14
4.2.1 Industry conferences and meetings 15
4.2.2 Awareness and networking events 16
4.3 Training for professionals 16
5 Conclusions 17
References 18

 © XLike consortium 2012 – 2014

List of Figures

Figure 1. Development version of WP5 visualization toolkit. 11
Figure 2: The XLike initial general flyer 14

List of Tables

Table 1. Summary of the showcase demonstrator. 12

 © XLike consortium 2012 – 2014

1 Introduction

Nowadays there is a real demand on harvesting data from different Web sources, especially from social media, to produce structured information which could be used later for different purposes as Business Intelligence, personalization and recommendation, searching, etc. This is mainly due to the fact that most of the information that is/has being created is unstructured and therefore it is sometimes difficult to find what you look for, and other times it is just not enough to have isolated information which does not provided any knowledge related with the context where it is happening.

This document provides in Section 2 a description of an industrial showcase where the main results of the XLike project would be applied and used in order to satisfy some of the detected needs of the industrial market. The presented showcase is related with social media monitoring and analysis for automatic news generation and contextualization. For this purpose, the complete set of tools provided by the different WPs are very valuable ranging from the natural language processing tools provided at WP2 (i.e. shallow and deep processing) to the analysis and visualization at WP5 (i.e. interfaces for accessing the services provided). It is crucial also to notice the importance of the WP1 which is going to be the one to provide the harvesting of web data for the showcase and it is also a major issue for the industrial partners because there is still a lack of structured data in the market.

The description of the requirements needed to successfully obtain a demonstratable showcase is presented in the section 3. For instance, a usual problem stems from the need to execute the crawling of web data, the natural language processing, the data analysis, and the visualization almost in real time and therefore a special attention must be paid to this issue. This is treated in the description of software/hardware general requirements in order to explain its importance. Although at this stage of the project it is still early to know exactly the amount of time needed by the different components of the showcase, we wanted to put it down for future discussions. In that section also a deeper description of the showcase is given indicating the different tools from the XLike Toolkit to be used for successful implementation of the showcase.

Other important point to be considered for this industrial showcase is how to reach and make the audience aware of the work done in the XLike project and its effort to provide successful tools in the context of Language Technologies, text mining, and cross-lingual information extraction. Regarding that point, a set of requirements to shape up the showcase and promotional material have been identified and they are presented in Section 4. Also we want to highlight the effort that needs to be done in order to fill the real gap between industrial and research conferences (which in many cases represents the real gap between academia and industry). For this purpose we have identified several industrial workshops at the major conferences on data mining and semantics in order to promote the work done in XLike with this industrial showcase.

Finally the conclusions are presented at Section 5.

© XLike consortium 2012 - 2014

2 Industrial showcase

The industrial showcase defined for the XLike project is to obtain a scalable platform to monitor different selected news streams and different social media (i.e. Twitter, Facebook, blogs, etc.) and to provide web services that make available the information related with the different events that are referred to in those media. Furthermore the events should be detected for the different languages that have been defined for the XLike project (English, German, Spanish, Catalan, Slovenian, and Chinese, possibly Hindi and Croatian) and the links between the different detected events in different languages should be discovered. The automatic generation of links to other blogs or articles, which are related with the detected events, also provides very relevant information for some specific industrial sectors or journalists or for feeding automatically contents in new Web portals.

We want to point out that in journalism a big effort is being made on early detection of events in order to provide a fast communication to the general public. Usually the newspapers and communication media have correspondents who report directly on the events or are permanently in contact with other sources of information which provide details on those events. The presented showcase belongs to the new industrial trend in data journalism helping to do this processing in an easier and faster way alleviating some of the traditional costs. This can open a door towards the industrial market.

The introduced platform will be responsible for executing the different defined pipelines for each language and also to provide an easy pluggable interface having a flexible framework in order to be adapted to other industrial showcases as needed. Therefore, the presented platform not only will provide the necessary services for detecting the events referred to in texts on different languages and find associated information, but will also provide the access to the complete set of generated data also via web services with possible different types of visualisation of detected data links.

Another characteristic of this industrial scenario is the large amount of data which is needed to be analyzed in real time in order to make it usable for the third parties. Therefore, a focus on how to increase the performance of the different offered functionalities and for the different languages or pipelines is mandatory. This task is challenging due to the continuous increase of the available data provided by the different news streams and social media, and the importance of having enough flexibility to adapt for future situations is highly desirable.

The proposed platform will also allow that the corpora, the data, and the toolkit incorporating the functionalities needed to successfully approach this showcase are going to be available beyond the end of the project not only for the scientific community but also for the industry as a set of services which will be easily accessible and pluggable into other systems or platforms as needed.

It is worth to highlight the importance of having a service oriented approach (SOA) which will provide also the opportunity of offering the functionalities provided by the project as SaaS (Software as a Service).

Some general requirements that this industrial showcase has to fulfil are:

- Creation of a scalable platform that allows to provide both, services offering the functionality needed for the above presented showcase and access to the generated structured information;
- Execution of the different language pipelines which provides the functionality needed to achieve the cross-lingual event detection as a part of the showcase prototype at M24;
- Real time performance of the platform in order to assure its functionality should be achievable, so that the streams from the monitored media sources can be updated;
- Presentation of the detected events and the associated generated information by using a visualization metaphor and allowing an easy and simple user computer interaction.

 © XLike consortium 2012 – 2014

3 Software Requirements and Components

In this section we cover the specific requirements needed to define the above presented scenario and the XLike components which cover the needed functionality in order to provide a final fully functional demo. The showcase requirements draw heavily on the *Early toolkit architecture* specification [1] and *Requirements for early prototype* [2].

3.1 Showcase software/architectural requirements

The general requirements presented in the introduction section can be specified deeper in the following set of architectural and software requirements:

- The platform needs to be extensible to a higher number of computers or to high performance distributed systems (i.e. Amazon Elastic Compute Cloud EC2[2]).
- The platform should be flexible enough to easily incorporate new services, or update the ones which already are deployed, by using a REST interface approach and alternatively SOAP.
- The different pipelines providing the basic functionality for each language have to be independent allowing its easy parallelization by using different threads or by using a map-reduce approach which is based on Hadoop[3,4] infrastructures.
- Both, the filtered data provided from the different resources and the data generated as result of the showcase have to be stored and accessible via services. For this purpose the use of MySQL and SPARQL endpoints (i.e. 4Store) are needed as part of the global platform.
- The generated/used data would be, whenever is possible, compliant with the META-SHARE standards for preserving them beyond the project
- The interaction with users will be by using simple GUI interfaces showing the report of the detected events at different languages and the associated news/articles.
- The interface will be also deployed as another component of the platform (i.e. Apache Wicket) and the visualization will make use of state of the art techniques such as HTML5.
- Specific event visualization templates could be used for easier interpretability.

The components will be as much independent as possible in order to avoid the cascading of failures and make the platform more robust.

3.2 Showcase – XLike required components

This section covers the components provided by the different WPs to execute the pipeline needed for the completion of the presented showcase. In the next section the description of the problem to be solved and the functionality classified by the WP that it belong to, are also presented.

Description of the problem: the problem to be address from a functional point of view is the cross-lingual linking of news articles, and detected and extraction of news events. Following the context of the trend in data journalism introduced above, a journalist may be interested in searching on the available news streams or main social media to see what is happening so s/he can find relevant new events to be communicated. Typically specific language event detection can be done by analyzing it independently but is not that easy to find the links with the same event on other languages in order to provide a richer context for it; for instance by aggregating articles in other languages related to the same event.

[2] http://aws.amazon.com/ec2/
[3] http://hadoop.apache.org/
[4] http://aws.amazon.com/elasticmapreduce/

© XLike consortium 2012 - 2014

Description of the possible solution: the M24 demo will solve the above problem by using the different components of the XLike Toolkit. More concretely the process will use:

1. Multilingual news stream data developed within WP1 will provide an uninterrupted large source of publicly available news data, to feed the public showcase,
2. Multilingual pre-processing services developed within WP2, including entity and relation extraction, which will be applied to articles from the news stream,
3. Semantic annotation developed within WP3, providing semantic context to extracted entities,
4. Cross-lingual document linking developed within WP4 will be used identify cross-lingual links between news stream articles,
5. Event detection prototype based on semantic templates developed within WP4, to represent detected events in a structured form.

As such, the showcase will make use of semantic and statistic approaches for contextualizing news stream.

The semantic approach will be used in order to create a set of event templates for the different languages which are based on the previously obtained semantic graphs. At these step also more complex patterns could be discovered as results of membership relations. It is also important to highlight the importance of using bilingual dictionaries constructed in WP3 in order to detect translational equivalents between languages and being able to find similar templates for the different languages. For this, WordNet and Cyc could be used jointly with machine translation to find the same different lexical groundings for the same concepts in the same and different languages.

The statistical approach will be used to find the cross-lingual mappings between the different languages and be able to cluster and connect similar articles across languages. This mapping can be done by using different techniques such as CCA or ESA, or other subspace learning method which can detect the relations between the different languages based on the words co-occurrence in comparable corpora. The advantage of this approach is lower reliance upon available high quality linguistic corpora (e.g. for learning deep parsers). We want to point out that, terms, keywords or named entities extraction can be used to provide a set of similar document in different languages by making a direct matching between these linguistic and knowledge units.

The showcase will consists of several workpackages or task level demonstrations, and a prototype demonstrating possible integration of the various components. For example, multilingual linguistic pipelines and semantic annotations tools will be wrapped in a textual enrichment online demo, similar to Enryhcer[5]. Similarly, cross-lingual document linking will be exposed as an online demonstration, showcasing cross-lingual similarity measure. There will also be an overall showcase or demonstrator, based on the visualizations developed within WP5 (Figure 1).

This description of the possible solution for the industrial showcase is summarized in Table 1.

[5] http://enrycher.ijs.si/

 © XLike consortium 2012 – 2014

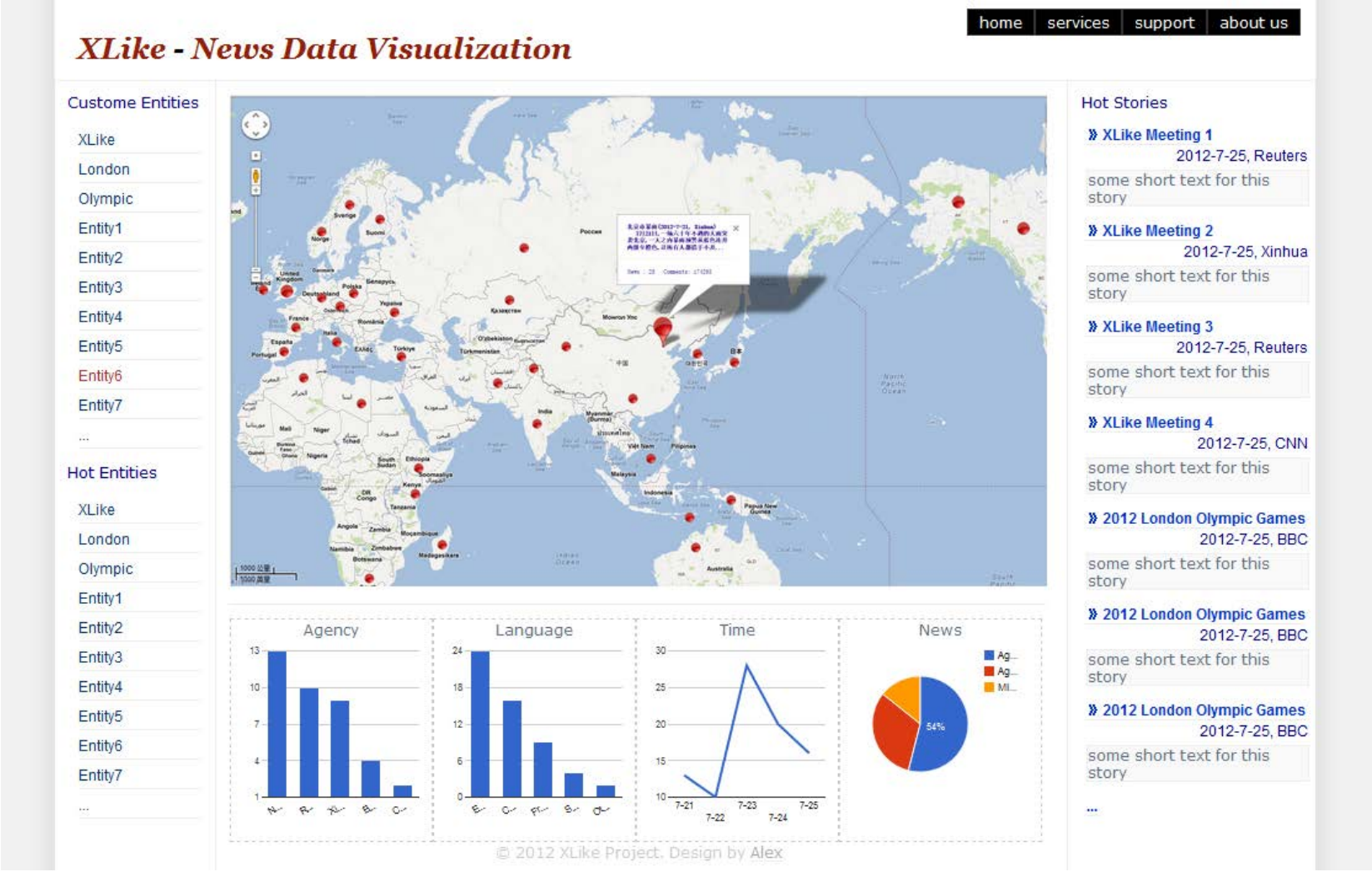

Figure 1. Development version of WP5 visualization toolkit.

Table 1. Summary of the showcase demonstrator.

	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 formal 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 formal sources, such as Twitter **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 for metaphors visualization of the detected events and the associated news at 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

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 will be focused on the following **target groups**:

- **Scientific and research community** – researchers in the areas of machine translation, corpus linguistics, computational linguistics and language technologies in general;
- **Industry and customers** – companies and professionals as potential users of XLike technologies interested in improving the quality of machine translation for under resourced languages and narrow domains;
- **General public** – interested in advancements in machine translation and language technologies;

Beside the general communication channels, such as web site, flyers, posters, t-shirts, etc., for each of this target groups different types of dissemination channels and activities are planned.

In this respect the target group we are interested in here, is the **Industry and customers** and for them we have introduced the industry outreach and education/training as our core dissemination channels.

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 spring and summer 2012 (e.g. LREC2012, EAMT2012, KTE2012, etc.), but it will be used also in the industry showcases as a general introduction to the XLike project.

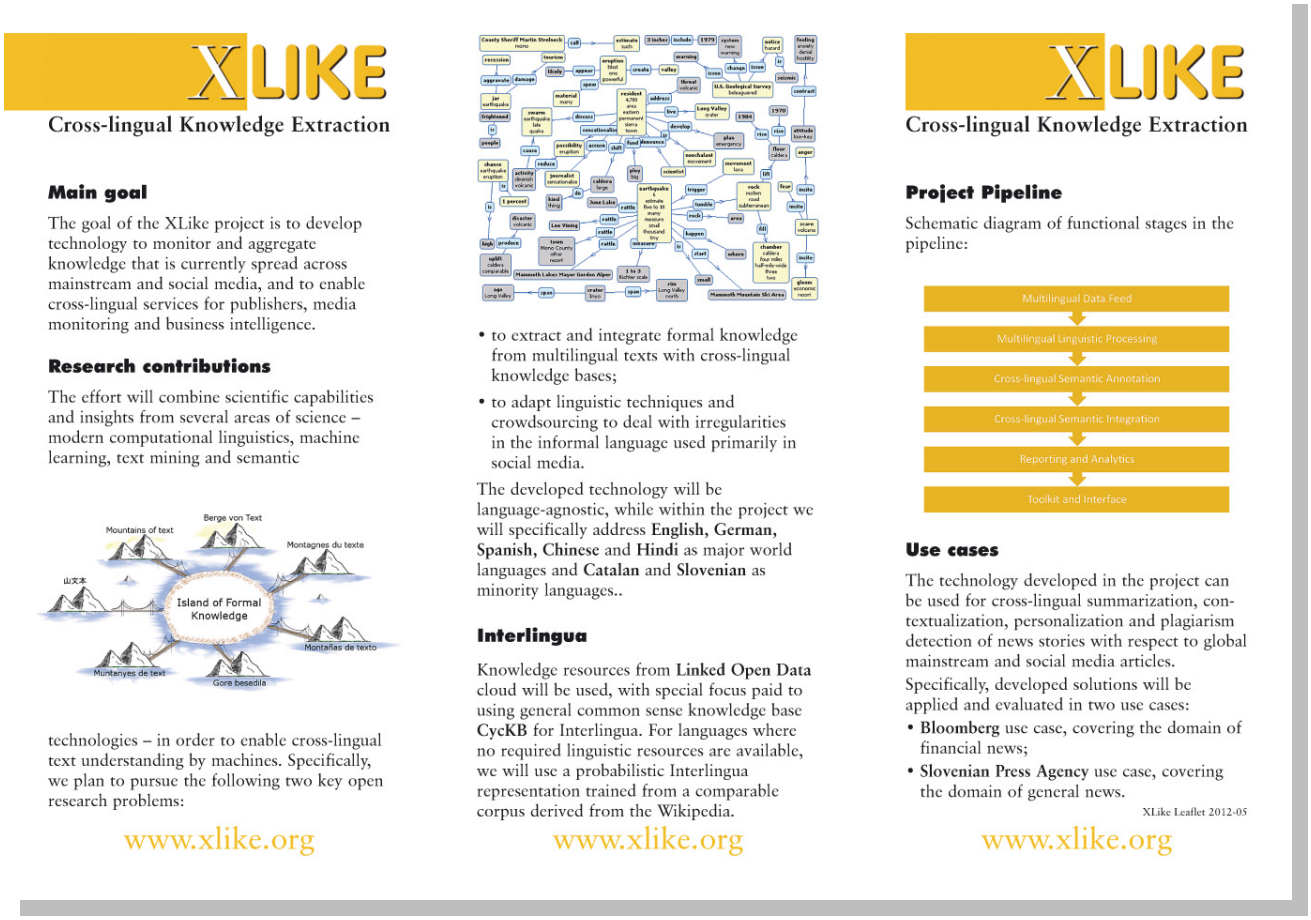

Figure 2: The XLike initial general flyer

For industrial showcase a limited number of flyers will be produced where that showcase will be presented. 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 final aim of these activities is that industry adopt the outcomes and to turn them into products and services. The use cases and the proof-of-concept prototypes are a valuable instrument to introduce XLike at industry conferences and business and technology congresses.

Actions and instruments foreseen for showcase demonstration:

- general view of the project (initial flyer)
- specific approaches to processing (described in special showcase flyer/leaflet)
    - linguistically motivated approaches (shallow & deep linguistic processing, semantic roles mapping to ontology relations, usage of Cyc-ontology/LOD, machine translation as aiding technology...)
    - statistically based approaches (detecting similar documents in different languages, CCA or ESA approaches, statistical machine translation as a mapping tool...)

 © XLike consortium 2012 – 2014

- specific application points of interest for the industry
    - harvesting web data (How to do it? How to make it available for industrial purposes? Solving IPR issues of web collected data.)
    - speed of processing (What kind of optimisations there are in XLike Toolkit?)
    - efficiency (How scalable background computing facilities, e.g. grid, are needed? What are typical computing scenarios and which computing resources have to be allocated for their processing?)
    - tuning the system for precision vs. 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?)

This 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.

The promotional requirements for industry showcase are:

- functioning prototype of XLike services
- assembled into a near-market level, i.e. near-industry strength processing pipelines
- applicable to the well-known text collections such as Wikipedia in all XLike languages, but stressing the general language-independence approach with several pluggable language-dependent modules
- applicable to the large streamlines of XLike industrial partners (BLO, NYT, STA...) in XLike languages
- applicable 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

The list below presents a number of relevant and recognized events and initiatives that will be attended having in mind the industry and customers target audience. Consortium members have attended some of these events regularly:

- **ESTC:** the yearly organized European Semantic Technology Conference is probably the most important event for industrial adopters and technology vendors in the fields of semantic computing in Europe.
- **SemTech:** SemTech is the American counterpart of ESTC, with over 1000 participants worldwide and a must for industrial dissemination of knowledge-driven technologies.
- **TextAnalytics Summit:** the annual summit is attended by numerous high-profile academic and industry representatives working in the text mining area.
- **I-**SEMANTICS and I-KNOW I-SEMANTICS is one of the largest conferences in Europe in the field of semantic systems and the Semantic Web. It is held concurrently with the I-KNOW Conference on Knowledge Management and Knowledge Technologies.
- Also, we will select a number of specialised, industry-oriented workshops that will be organised during the large scientific conferences to present XLike to industrial target audience. In this way we expect to fill the gap between data-mining community and industry that is still present (i.e. industrial/government track at KDD or ICDM conferences).

In this respect XLike will expect from the industrial partners to take important role in dissemination of the project's results and achievements in their own context at these events. At such industrial events prepared

© XLike consortium 2012 - 2014

manned public showcases and demos will be used to maximise the disseminating efficiency and provide the possibility for direct Q&As.

4.2.2 Awareness and networking events

We consider as very important different awareness rising and events oriented towards networking between research and industry where the project partners can disseminate information about XLike. The list below is by no means exhaustive, but it reflects the most important events expected within the project duration:

- LT Days
- LT Innovate
- ICT Days
- ...

These events are usually organised in the form of exhibitions with booths where XLike will participate with different sets of dissemination instruments. At such events prepared public showcases and demos will also be used. This approach is also expected after M24 when the prototypical solutions will be developed to a near-to-market level and will be tested and then promoted in public settings.

4.3 Training for professionals

Two of our audience target groups are also expected to attend different training activities: academic education and training for professionals. In this deliverable the second type of training activities is relevant. It targets companies and similar corporate and public bodies. The XLike partners considered in this activity are mainly focused on informing potential customers and early adopters of XLike technologies about the possibilities and advantages of a cross-lingual knowledge extraction empowered by the usage of XLike sophisticated mining methods, lightweight semantics, and powerful natural language processing.

The industrial showcase will build a good part of the training for professional activities, since it is expected that that audience target group will come mainly from industrials interested in XLike technologies.

XLike will provide tutorials and webinars on the usage of the XLike Toolkit, and learning resources in form of video lectures (advertised through videolectures.net at JSI) and presentations (hosted on the project portal, advertised through the REASE repository at STI International of which KIT and JSI are members).

 © XLike consortium 2012 – 2014

5 Conclusions

This document has presented the description and first specification of the industrial showcase which is going to be accomplished for the functional demo at M24. The showcase main goal is to detect at news streams and different social media a set of events for different languages which have been identified for the XLike project (English, German, Spanish, Catalan, Slovenian, and Chinese, possibly Hindi and Croatian) and provide an automatic generation of links to other blog or articles which are related with the detected events. This showcase seems to be very relevant for journalists or for feeding automatically contents in new Web portals.

A deeper understanding of the showcase has been also presented and described making special emphasis in the needed components and the pipeline structure that XLike project will need to provide. This is the first deliverable of two where we have defined the initial specifications for the showcase and will be followed by a fully functional demo at D8.2.3 XLike Showcase.

A set of promotional material for industrial awareness and outreach has also been defined and already some of them are available for the project. The diffusion of the results of the showcase at industrial conferences and industrial workshops of major data mining and semantic conferences has been also proposed as an activity for filling the gap between industry and academia.

Most of the work to be done in the presented context has to do with making the necessary activities for having the functional demo for M24 and presenting the main outcomes to the industrial community for industrial awareness.

© XLike consortium 2012 - 2014

References

[1] D6.1.1 – Early toolkit architecture specification

[2] D1.2.1 – Requirements for early prototype

 © XLike consortium 2012 – 2014