XLike

Deliverable D1.2.1

Requirements for early prototype

Editor:	Achim Rettinger, KIT
Authors:	Achim Rettinger, KIT; Lei Zhang, KIT; Blaž Fortuna, JSI; Peter Penko, STA; Aleš Pečnik, STA; Pat Moore, Bloomberg
Deliverable Nature:	R
Dissemination Level: (Confidentiality)	PU
Contractual Delivery Date:	M3
Actual Delivery Date:	M3
Suggested Readers:	All project partners
Version:	1.0
Keywords:	use case requirements; early prototype

Disclaimer

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

All XLike consortium parties have agreed to full publication of this document.

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:	WP1 – Definition and Data Provision
Document Title:	D1.2.1 – Requirements for early prototypes
Editor (Name, Affiliation)	Achim Rettinger, KIT
Work package Leader (Name, affiliation)	Blaž Fortuna, JSI
Estimation of PM spent on the deliverable:	7

Copyright notice

Executive Summary

This document presents description of the functional requirements for the early prototype of the Xlike project. It also provides a review of case studies on financial news and general news with the aim on how to apply and integrate key research contributions of the project into production pipeline of two news agencies.

The goal of this task T1.2 is to gather and survey the requirements for the XLike system based on the case studies. The main outcome of the task will be three requirements documents on (a) how to integrate solution for cross-lingual information linking and knowledge extraction with the business processes within the companies, (b) to define a set of possible services built on the top of the XLike technology which will be used within case studies and (c) to position supported languages in the context of case studies.

This deliverable is the first outcome of the task T1.2, which provides functional specifications for the early prototype, using the pre-existing technologies identified in the task T1.1 and the technology developed in the first year of the project. In order to keep pace with the technology developments within the project, the requirements will be revised at the beginning of each year based on the feedback from the validation process and this document will serve as an initial basis for future updates.

Table of Contents

List of Figures

List of Tables

Abbreviations

STA   Slovenian Press Agency

BLP   Bloomberg

1   Introduction

The purpose of this document is to gather and survey the functional requirements for the early prototype of the XLike project. We will review the use cases on financial news and general news, and jointly with the existing technologies from all partners of the project, they motivate a number of requirements for the development of XLike technology.

1.1   Use Cases

There are two use cases that will provide feedback end evaluation for major languages (English, Spanish, German, and Chinese), as well as minority languages (Slovene). They will be run by the following use case partners:

- *Bloomberg (BLP)* [1] – **financial news source** of information for businesses and professionals. Bloomberg combines analytic, data, news, display and distribution capabilities to deliver critical information via their service and multimedia platforms. Bloomberg's media services cover the world with more than 2,200 news and multimedia professionals at 146 bureaus in 72 countries.

- *Slovenian Press Agency (STA)* [2] – national press agency covering domestic and international events. STA offers **general news service** in Slovenian and daily English service and is the only provider of daily news in English for the expatriate community in Slovenia and for English-speaking readers abroad.

The use cases have been carefully selected in order to demonstrate the advantages of the project results, in respect to the three issues from the vision statement. Both case studies are news agencies. The first, Bloomberg, is mainly focused on fast and accurate delivery of financial and business news in English. The second, Slovenian Press Agency, is focused on delivering general news, in English and Slovene with focus on events happening in Slovenia or related to Slovenia.

The project will cover four applications which different stake-holders will benefit from within the use cases:

- **Cross-lingual Summarization** – Editors in the use cases need a good overview of the global developments in the areas and topics they are covering. The XLike project will develop applications, which enable cross-lingual summarization of real-time news streams coming from global and local mainstream media as well as from hyper-local social media. This will provide the editors with an overview of what are the most current topics, entities and events covered in the global media, how the information is spreading across the media sources and languages, and identify significant trends.

- **Cross-lingual Contextualization** – Understanding news article depends on the background knowledge of the user and the context presented by the author of the article. The XLike project will develop applications, which enable on-the-fly contextualization of articles. This will be achieved by displaying the provenance of information within the article (linking to past articles across languages with parts of content in the article at hand), diversity of views on the topic discussed in the article (displaying related articles across languages with different points of view) and visualization of related stories across languages.

- **Cross-lingual Personalization** – The trend in news delivery is towards personalization, recommending and displaying the articles that are relevant to the user at the context. The XLike project will develop applications based on the current methods for news story recommendation (available through the collaboration between Bloomberg) which allow for cross-lingual recommendation systems using document linking and annotation with cross-lingual knowledge bases.

- **Cross-lingual Plagiarism Detection** – The XLike project will develop applications enable monitoring for unauthorized reuse of copyrighted content on the Web.

1.2   Developed Applications within the Use Cases

To illustrate the possibilities offered by the XLike technology, we detail the set of applications which will be developed within the use cases.

1.1.1   Cross-lingual Summarization

Description of the problem – Covering a news topic requires good understanding of the topic and wide coverage of all the related news being produced about the topic. Rise of social media only increased the amount of data produced on almost any topic and separating signal from noise is becoming an even bigger challenge. Traditional ways of handling this is by relying on several reliable sources and manual filtering of information. This introduces latency which is not acceptable in domains such as financial news.

Description of possible solutions – This application will solve the above problem using components for information visualization (T5.2) and trend detection (T5.3). Three components will be used to provide the editors with an overview of mainstream and social media by summarizing and visualizing, across news sources, languages and time:

- Current topics, entities, and events

- Information spreading across global media

- Identified significant trends

The solution will provide the editors with better understanding of recent developments in the news topics they cover and will indicate demand on the publishing market – what the publisher should write about because it is relevant to their audience and not yet or poorly covered.

Description of the evaluation – There are several hard (directly measurable) and soft (subjective perception of the involved parties) metrics that can be used to evaluate the usefulness and effectiveness of a system:

- Satisfaction of editors based on the overview provided (Does it help or confuse the users?)

- Latency of breaking news detection using large media stream

1.1.2   Cross-lingual Contextualization

Description of the problem – Understanding a news article depends on the background knowledge of the user and the context presented by the author of the article. However, typical articles would have several points of view and a typical user does not have time or know how to analyse and discover additional source relevant to the article at hand.

An article can also have consequences, which are not directly obvious to the reader, but might be important. This is especially the case in financial news, where an event in one market can have an effect on other markets, related by some chain of causality.

Description of possible solutions – This application will solve the above problem using components for information flow (T5.2), reporting bias and sentiment analysis (T5.1) and complex event detection (T5.3). These components will provide the users with an overview of mainstream and social media, which is related to a particular news story, by:

- Displaying the provenance of the information displayed in the article across media sources, languages and time

- Diversity of views on the topic and subject of the article across more sources and languages

- Visualizing other related or similar stories Solutions from complex event detection will be used to detect causality chains and alert users of the potential effects for a particular article.

Description of the evaluation – There are several hard (directly measurable) and soft (subjective perception of the involved parties) metrics that can be used to evaluate the usefulness and effectiveness of a system on known closed-world scenarios, where we know the ground-truth:

- Evaluation of provenance tracking

- Using discovered bias in reporting for provide an overview of main views for particular topics

- Number and accuracy of detected complex events

1.1.3   Cross-lingual Personalization

Description of the problem – The trend in news delivery on the Internet is towards personalization, recommending and displaying the articles that are relevant to the user at the given time and context. Typically, news personalization platforms rely on language dependent features and cannot be generalized across languages. Hence, even if a typical user is capable of understanding more languages, the limitation of current systems prohibits reliable cross-lingual recommendation service.

Description of possible solutions – This application will solve the above problem using components for cross-lingual linking of articles based on topics (T4.1) and annotation with cross-lingual knowledge resources (T3.1) to bridge the gap between languages present in the modern recommendation systems (available within consortium through collaboration between JSI and Bloomberg). The application will observe and store users' profiles and, based on the profiles, deliver up-to-date recommendation using the API developed in T6.3. An additional service offered by the application will be a cross-lingual semantic search engine, which will rely on cross-lingual document linking and cross-lingual knowledge resource to match the user query against multilingual news article corpora.

Description of the evaluation – There are several hard (directly measurable) and soft (subjective perception of the involved parties) metrics that can be used to evaluate the usefulness and effectiveness of a system:

- Success rate of recommendations (how likely are users to click on the recommendation)

- Satisfaction of users given the multilingual recommendations

- Satisfaction of users given the multilingual search results

1.1.4   Cross-lingual Plagiarism Detection

Description of the problem – Slovenian Press Agency covers domestic and international events by producing and selling copyrighted content. However, protection and tracking of copyrighted material on the Web is an open problem especially when the material involved is translated in the process.

Description of possible solutions – This application will solve the above problem using the components for cross-lingual document linking and information flow to detect articles which have significant overlap with the source article.

Description of the evaluation – There are several hard (directly measurable) and soft (subjective perception of the involved parties) metrics that can be used to evaluate the usefulness and effectiveness of a system on known closed-world scenarios, where we know the ground-truth:

- Number and accuracy of detected copies with the source language

- Number and accuracy of detected copies across languages

- Low cost per plagiarism check for one article

2   Requirements for the Early Prototype

In this section, we will review the two mentioned use cases. Based on the detailed analysis of case studies needs and the pre-existing technologies available within the XLike Consortium, we derive the functional requirements for the early prototype that are essential to the use cases.

2.1   Prototype of Bloomberg Use Case

Bloomberg's business is the delivery of financial information. The core of their business is based on Bloomberg Terminals, a specialized platform for financial professionals. Besides this, they also maintain a more mainstream oriented news portal at Bloomberg.com. The Bloomberg use case in XLike will focus on the website, by evaluating techniques for cross-lingual integration of news articles.

2.1.1   Entity Tracking

The Bloomberg.com website maintains a section on market information. As part of the section, each major company has a dedicated page, listing core statistics, as shown on Figure 1. The company profile page contains a list of latest news articles, related to the company, pooled from the rest of Bloomberg.com. This works well for major or US companies, for which enough content is produced. However, the list might maintain outdated articles for smaller companies or companies from other parts of the world. For example, the list from Figure 1 contains articles, which were 4 days old when the screenshot was taken.
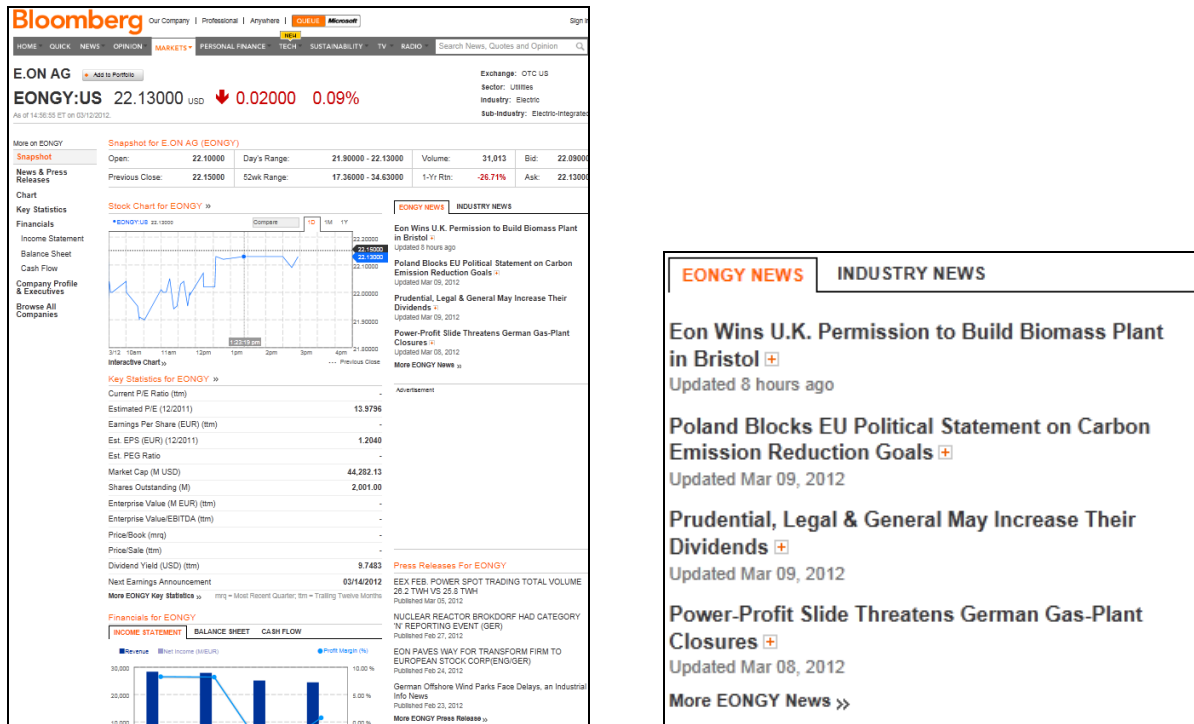

Figure 1. Company profile for E.ON a.g. (left) and zoom-in to latest company news (right).

The first task in the Bloomberg use case is to generate a more up-to-date list of relevant company news, preferably from their home markets. To follow the example from Figure 1, on the same day, there were more than 1000 new German articles with keyword "EON" on Google News. The task can be roughly defined in two steps as (1) detect mentions of the entity in the multi-lingual news stream and (2) determine which four are most suitable to be displayed on the company news profile page. The first step requires entity extraction from multi-lingual stream, while the second step requires integration and summarization across all languages with relevant articles.

Formally, the task is defined as follows:

1.   For each entity e_i from a set of entities $E = \{e_i, ..., e_n\}$, identify articles $A_{e_i} = \{a_j \mid e_i \in a_j\}$.

2. For a set of articles A_{e_i}, identify a subset of k articles best summarizing the set for some small value of k. Summarization should take into account magnitude and diversity events reported in the selected articles compared to complete set A_{e_i}.

Table 1. Entity Tracking in Bloomberg Use Case.

	Entity Tracking
Application	Cross-lingual contextualization and Cross-lingual summarization
Input	a) A list of entities b) Mainstream news stream
Output	a) A list of recent relevant articles for each entity b) A subset of articles best summarizing relevant articles for each entity
Languages	XLike languages
Related tasks in Y1	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news services for article tracking **T2.1 – Shallow linguistic processing** of formal language used for entity extraction from mainstream articles **T3.1 – Approximate text annotation** with cross-lingual semantic repositories used for entity disambiguation **T4.1 – Statistical cross-lingual document linking** used for cross-lingual summarization **T5.2 – Information visualization** used for visualization of matching articles
Evaluation	a) Number of tracked articles b) Relevance of summarization articles (focused user study)

2.1.2 Related or Relevant Articles

Bloomberg.com provides personalized list of suggested articles along each article. The list is assembled from the recent Bloomberg articles and custom fitted for the specific user, based on his/her history. The second task in Bloomberg use case is to extend the suggested articles by including external mainstream sources across more languages.

Table 2. Related or Relevant Articles in Bloomberg Use Case.

	Related or Relevant Articles
Application	Cross-lingual contextualization and Cross-lingual recommendation
Input	a) Mainstream news stream b) Social media stream
Output	a) A set of matching articles b) Summarization and Visualization of matching articles
Languages	XLike languages
Related tasks in Y1	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news services for article tracking **T4.1 – Statistical cross-lingual document linking** used for article linking and user recommendation **T5.2 – Information visualization** used for visualization of matching articles
Evaluation	a) Accuracy of identified articles matching Bloomberg.com articles b) Relevancy of recommended articles (focused user study)

Formally, the task is defined as follows. Given a user u, with visit history $H(u) = \{a_1, ..., a_n\}$, identify relevant recent articles from multi-lingual news stream. All articles from $H(u)$ are in English and published

by Bloomberg. Assembling a relevant recent articles list requires cross-lingual integration with Bloomberg.com articles.

2.2 Prototype of STA Use Case

STA publishes article in Slovene and English. Primary market for Slovene articles is Slovenia, which either republish the articles or use their content as input to their own articles. The primary target for English articles is foreign news agencies (e.g. Xinhua) or news outlets (e.g. Bloomberg). Some of foreign agencies also use Slovene articles.

2.2.1 Article Tracking

At the moment, STA does not have means to track the republishing of its articles. There are two business cases for why such a tracking is important. First, the main income of the agency is licensing its content, and publish unlicensed material requires their attention. Second, knowing which articles are republished by their subscribers helps the agency to better understand their market, and to provide better coverage for the events relevant for them.

More formally, article tracking should detect the following modification operations to article a_1:

1. Article a_2 is a near-copy of the other: $a_1 \approx a_2$
2. Article a_2 is a translation of the original article: $T(a_1) \approx a_2$
3. Article a_2 includes article a_1: $a_1 \subset a_2$
4. Article a_2 includes part of article a_1: $P(a_1) \subset a_2$

Ideally, any above operations, or combination of them, should be detectable. Given the technology available within Y1, work on the first two operations will be emphasized.

Table 3. Article Tracking in STA Use Case.

	Article Tracking
Application	Cross-lingual contextualization and Cross-lingual plagiarism detection
Input	a) STA article stream b) Mainstream news stream
Output	For each STA article, a set of matching articles from the mainstream news stream
Languages	a) STA article are in Slovene and English b) Focus on XLike languages
Related tasks in Y1	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news services for article tracking **T4.1 – Statistical cross-lingual document linking** core technique used for the development of this component **T5.2 – Information visualization** used for visualization of matching articles
Evaluation	a) Number of tracked articles within the source language b) Number of tracked articles across languages

2.2.2 Topic and Entity Tracking

STA covers topics related to Slovenia or Slovenian entities (E.g. companies, athletes). As such, tracking relevant news is an important part of editors' daily routine. Technologies developed within XLike project can improve this process by providing tools for detecting relevant articles across languages and media (mainstream, social media).

Formally, topic or entity tracking can be seen as a filter applied to a stream of articles. An article is retained by the filter if it matches the topic, or is related to the entity. Topics can be defined as a standard classification task, with articles on the input and set of matching topics on the output. Entities can be detected using named-entity extractors.

For popular topics or entities, the filter can retain a large amount of articles. The information contained within these articles can be visualized or summarized to help the editors in skimming through the content, to identify relevant events.

Table 4. Topic and Entity Tracking in STA Use Case.

	Topic and Entity Tracking
Application	Cross-lingual contextualization and Cross-lingual summarization
Input	a) Mainstream news stream b) Social media stream
Output	a) A set of matching articles b) Summarization and Visualization of matching articles
Languages	XLike languages
Related tasks in Y1	**T1.3 – Data infrastructure** must provide sufficient corpora of existing articles for experimentation and sufficient coverage of relevant mainstream news services for article tracking **T2.1 – Shallow linguistic processing** of formal language used for entity extraction from mainstream articles **T3.1 – Approximate text annotation** with cross-lingual semantic repositories used for entity disambiguation **T4.1 – Statistical cross-lingual document linking** used for topic tracking **T5.2 – Information visualization** used for visualization of matching articles
Evaluation	a) Precision of detected articles b) Recall of detected articles on controlled dataset c) Effectiveness of summarization and visualization (user questionnaire)

2.3   Functional Requirements for Early Prototype

According to the detailed analysis of both use cases, the functional requirements for the early prototype described below must be implemented to allow the users to perform each use case. Each requirement includes a short description and is referred to the corresponding tasks in the project.

Table 5. Functional Requirement of Shallow Linguistic Processing.

	Shallow Linguistic Processing
Description	To prepare and develop tools for shallow linguistic processing of formal language corpora.
Input	A sentence, a document or a set of documents in any XLike language
Output	a) The language given content is in b) Sentences, words, tokens, POS tags, ect. c) Named entities
Languages	XLike languages
Motivation	Entity tracking (BLP), Topic and entity tracking (STA)
Related task	**T2.1 – Shallow linguistic processing of formal language**.
Evaluation	a) Accuracy of the resulting lexical items b) Performance of the shallow linguistic processing

Functional requirement 1: Shallow linguistic processing

Regarding this functional requirement, we need to develop tools for shallow processing across all required languages. These tools are supposed to include language detection, tokenization, stemming and lemmatization, part-of-speech tagging, bag-of-words generation and named entity extraction. Named entity extraction can be done with the help of external multilingual resources such as multilingual Wikipedia corpus and Linked-Open-Data providing large coverage of relevant entities across many languages.

Functional requirement 2: Early text annotation

For the functional requirement of early text annotation, we need to develop tools to perform lightweight approximate annotation between text, represented in structures and semi-structured forms as result of WP2, and selection of relevant knowledge resources such as Wikipedia, Linked-Open-Data, Cyc. Machine translation techniques can be used to provide cost-effective cross-lingual lexical grounding of resources for the 6 languages targeted by the project. The result of this task will be knowledge resources with cross-lingual lexical grounding and tools for probabilistic annotation of documents which will need further disambiguation.

Table 6. Functional Requirement of Early text annotation.

	Early text annotation
Description	To annotate documents with cross-lingual language and knowledge resources
Input	a) One or a set of documents in any XLike language b) Relevant knowledge resources, such as Wikipedia, Linked-Open-Data and Cyc
Output	a) Translation of existing lexical groundings from knowledge resources to all required languages b) Annotations of words in the document with one or more concepts from knowledge resources based on their lexical information
Languages	XLike languages
Motivation	Entity tracking (BLP), Topic and entity tracking (STA)
Related task	**T3.1 – Approximate text annotation with cross-lingual semantic repositories**.
Evaluation	a) Precision of the text annotation b) Recall of the text annotation c) Performance of the translation and the text annotation

Table 7. Functional Requirement of Cross-lingual document linking.

	Cross-lingual document linking
Description	To develop techniques for cross-lingual document linking based on statistical models and cross-lingual knowledge resources
Input	One document in any XLike language
Output	A set of documents based on their similarity to the input document across languages
Languages	XLike languages
Motivation	Entity tracking (BLP), Related or relevant articles (BLP), Article tracking (STA), Topic and entity tracking (STA)
Related task	**T4.1 – Statistical cross-lingual document linking**.
Evaluation	a) Precision of linked documents b) Recall of linked documents c) Performance of the model training and document linking

Functional requirement 3: Cross-lingual document linking

For this functional requirement of cross-lingual document linking, we need to create a statistical mapping model for detecting topic similarity across different languages by using techniques from cross-lingual information retrieval. Comparable corpora such as Wikipedia with cross-lingual links can be used to train such a model. We need to train statistical mappings for all languages required by case studies, plus additional languages with sufficient support in available comparable corpora.

Functional requirement 4: Early information visualization

With respect to the functional requirement of early information visualization, we need to show real-time information spreading across the globe for the events and news using some of the existing text visualization and network visualization techniques. The emphasis will be given to innovative approaches on how to visualize temporal and highly structured data, which will be available as result of WP2, WP3 and WP4. The outcome will be a set of components which will be used in the X-LIKE Toolkit (WP6).

Table 8. Functional Requirement of Early information visualization.

	Early information visualization
Description	To employ techniques for text and network visualizations for real-time cross-lingual streams to show visual summary of information dynamics across sources, languages and time
Input	Documents or corpus representations from the previous stage
Output	Visualization of the input documents or corpus representations
Languages	XLike languages
Motivation	Entity tracking (BLP), Related or relevant articles (BLP), Article tracking (STA), Topic and entity tracking (STA)
Related task	**T5.2 – Information visualization.**
Evaluation	Effectiveness of visualization (user questionnaire)

3   Conclusions

This deliverable provides the functional specifications for the early prototype of the XLike project based on a critical review of the two use cases: **Bloomberg** and **Slovenian Press Agency**. With respect to the main result of the task T1.2, jointly with this deliverable, there are three requirements documents, one at the beginning of each year, providing detailed analysis of case studies needs and service opportunities using innovations and solutions from the project for the prototype to be developed in the respective year.

This document is the first output of the task T1.2 and will be further used as source of information for the M15 deliverable **D1.2.2 – Requirements for demonstrator**, and the M26 deliverable **D1.2.3 – Requirements for fully functional prototype**.

References

[1]   Bloomberg (http://www.bloomberg.com)

[2]   Slovenian Press Agency (http://www.sta.si)