xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D7.3.1

Early Prototype and Validation Report MONITOR

Editor:	Anna Björk Nikulásdóttir, VICO
Author(s):	Anna Björk Nikulásdóttir, VICO
Deliverable Nature:	Prototype (P)
Dissemination Level:	Confidential (CO)
Contractual Delivery Date:	M12 – 31 October 2014
Actual Delivery Date:	M12 – 31 October 2014
Suggested Readers:	All project partners, users of (social) media monitoring systems
Version:	1.0
Keywords:	VICO, Use Case, Monitoring

Disclaimer

Full Project Title:	xLiMe – crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP7 Use Cases and Evaluation
Document Title:	D7.3.1 – Early prototype and validation report MONITOR
Editor:	Anna Björk Nikulásdóttir, VICO
Work Package Leader:	Ronald Denaux, ISOCO

Executive Summary

This document describes the development of the Year One Early Prototype of the xLiMe Use Case "Monitor", performed by the xLiMe partner VICO Research and Consulting GmbH [1].

The use case "Monitor" can be summarized as **cross-lingual, cross-media brand monitoring and topic analysis**. The aim is to monitor brands and opinions across languages and media over time, extending VICO's social media monitoring system. This includes the integration of data sources and technical resources provided and developed within the xLiMe project as well as the implementation of additional xLiMe specific features in VICO's system.

In the early prototype version introduced in this document the emphasis was on the integration of data sources from the xLiMe partners, including the continuously delivered social media data from VICO. The xLiMe infrastructure offers possibilities for the partners to consume these data, whereas the adaption of the data to VICO's system was implemented locally. Furthermore, some adjustments of VICO's system were needed to be able to deal with all xLiMe data.

The system now can include TV speech-to-text transcripts [2] and news articles delivered by xLiMe partner JSI [3] in its dashboard and analysis. Currently the inspection of the data is performed using keyword search but preparation for an entity based search is already in process. In an entity based search annotations with entities from a knowledge base are consulted whereas a keyword based search looks for a direct match in the text.

The evaluation process of the early prototype is already planned and will be performed by analyst specialists at VICO. The analysts will evaluate the tool from the user's point of view in order to estimate the impact the additional data sources have on the results compared to the restriction to an analysis based solely on social media.

Table of Contents

List of Figures

Abbreviations

KIT	Karlsruhe Institute of Technology, Karlsruhe, Germany
JSI	Institute "Jožef Stefan", Ljubljana, Slovenia
OCR	Optical Character Recognition
RDF	Resource Description Framework

1 Introduction

This document describes the early prototype of the VICO use case "Monitor". The goal of the use case is to monitor brands and opinions across languages and media types.

In this section we will briefly introduce VICO's social media monitoring system, with emphasis on features related to the xLiMe project. Then a general summary of the use case is given and the constraints for the early prototype developed during year one (Y1) of the project are introduced. The following sections discuss the early prototype in more detail: Section two describes the data sources and data processing, section three shows the extended VICO dashboard and section four covers an actual use case demonstration. Before concluding and describing future work in section six, section five defines the evaluation planned for the Y1 prototype.

1.1 VICO's Social Media Monitoring System

VICO harvests large amounts of data from social media. The sources include large social networks like Twitter, Facebook, Google+, and YouTube, but also a broad spectrum of forums, blogs, review sites, and Q&A portals. A number of news portals and other websites are also included in the list of analysed data.

The system already covers over 40 languages. The search queries are modelled separately for each language relevant for the customer. Queries can get very complex, aiming at getting the best results regarding the customer's subject of interest.

The monitoring results are presented to the user in a dashboard. A number of analysing metrics are available, including sentiment, tag clouds, volume, importance of sources, etc. Additionally, over 50 filtering options allow a specific view on the data and analysis results.

We can summarize the features of the current monitoring system which have a direct relation to aspects of the xLiMe project:

- The data predominantly consist of texts from social media with news portals as an additional source of data.
- The system can handle texts in more than 40 languages.
- Keyword based search queries modelled for each language, high complexity for best results.
- Sentiment analysis.
- Tag clouds based on parts-of-speech of words or on Twitter hashtags.

1.2 The VICO xLiMe Use Case "Monitor"

For VICO's xLiMe use case the current monitoring system is being extended using data and technologies provided and developed within the xLiMe project. Thanks to the flexible HTML5-based VICO dashboard, it is possible to test new features and data sources in line with the rapid development work flow of the xLiMe project [4].

The use case is defined as follows:

Cross-lingual, cross-media brand monitoring and behaviour analysis: *Monitoring of brands and public opinion, across languages, content types and content generation types to enable seamless trend prediction and issue management for organizations beyond local markets.* [5]

VICO already monitors brands and opinions across languages using language specific query modelling. The planned search by entities and events, making use of the streaming interface of the xLiMe data processing

infrastructure [6] and [7], will move the search more towards a true cross-lingual search. This means each language will not have to be handled separately.

Data sources from the xLiMe partners JSI and ZATTOO [8], i.e. news articles and TV data, will be used along with the social media sources already in use in the current system, aiming at cross-media monitoring and analysis.

The TV data is processed on different levels. There is a speech-to-text transcription [2], OCR from video [9], and annotations based on content recognition in the video stream [10]. The content annotations component will be used in the Monitor use case to monitor occurrences of brand logos in TV shows.

The inclusion of TV-sources does not only contribute to a more diverse analysis based on different media types for all customers, but also has the potential to reduce manual work of customers who have employees watching TV to monitor subjects of interest. Embedding TV channels and their different processing levels (speech-to-text, logo detection, etc.) in the social media monitoring would be a unique feature of VICO's system on the market.

1.2.1 Y1 Early Prototype

In the early prototype of the VICO use case the emphasis is on integrating the different source types above cross-lingual search. This is due to the general state of development in the xLiMe project. Data are being pushed into a central data queue [6] whereas the streaming interface will be published in year two of the project, enabling e.g. cross-lingual entity search [6]. The preparation for entity search is already in progress but the early prototype will only be based on keyword search on different source types, monitoring brands and subjects of interest over time.

2 Data Sources and Processing

Three participants of the xLiMe project provide data: Zattoo (TV data), JSI (textual news data from online news sites) and VICO (textual social media data). In order to be able to use the data in the VICO system, it has to be fetched from the xLiMe data storage and then converted into an internal format.

2.1 Data from VICO

VICO provides social media data to the xLiMe data storage. VICO's internal data queue is filtered by pre-defined keywords, languages and data sources to limit the data volume to push to the xLiMe queue. Before pushing the data the documents are annotated using a Wikipedia annotator developed at KIT [11] and then converted from VICO's custom document model to the xLiMe specific RDF-format [12].

2.2 Sources used in Y1 Early Prototype

All sources available by M11 of the project are used in the early prototype. These are: news articles from JSI, VICO's social-media data and TV speech-to-text data. The inclusion of TV-OCR data and TV content annotation data will be left for year two, as the data will become available.

2.3 Data Processing

As described in [6] and [7] the data providers push their data into a common data queue. Each kind of data stream gets associated with a named topic, e.g. the TV speech-to-text data is currently pushed to the topic *zattoo-asr*. Each partner can already consume the topics directly from the queue. During year two of the project more advanced consuming processes will be available. These will include making use of two components defined in [7], the streaming data provider and the historical data provider. For the Y1 prototype the VICO data processing unit consumes the topics from the data queue associated with the data sources described in section 2.2: *tv-metadata, zattoo-asr, jsi-newsfeed* and *socialmedia*.

The consumed documents then need to be converted into the document format suitable for further processing by VICO's monitoring system. As each data source has a slightly different RDF-format and uses different RDF-predicates, a separated conversion procedure had to be implemented for each source (see a detailed discussion of the xLiMe data formats in [12]). The converted documents are then stored in an index which is consulted by VICO's text analysis units like tagger (preparing tag clouds) and sentiment analysis. Since data from TV channels represent a new source type for the system, the data model and further processing units needed to be extended to deal with this new type.

3 The Dashboard

As stated in the Introduction, VICO's HTML5-based dashboard is very flexible, which makes the development of special extensions for the xLiMe use case a feasible task. For the Y1 early prototype the main task was to add the new source type TV and to connect it to the underlying data model. The dashboard offers the possibility to filter result data by source types and thus to perform an analysis on certain source types like e.g. TV data and news articles only. Figure 1 shows an example layout of the dashboard including the new data source type "TV".

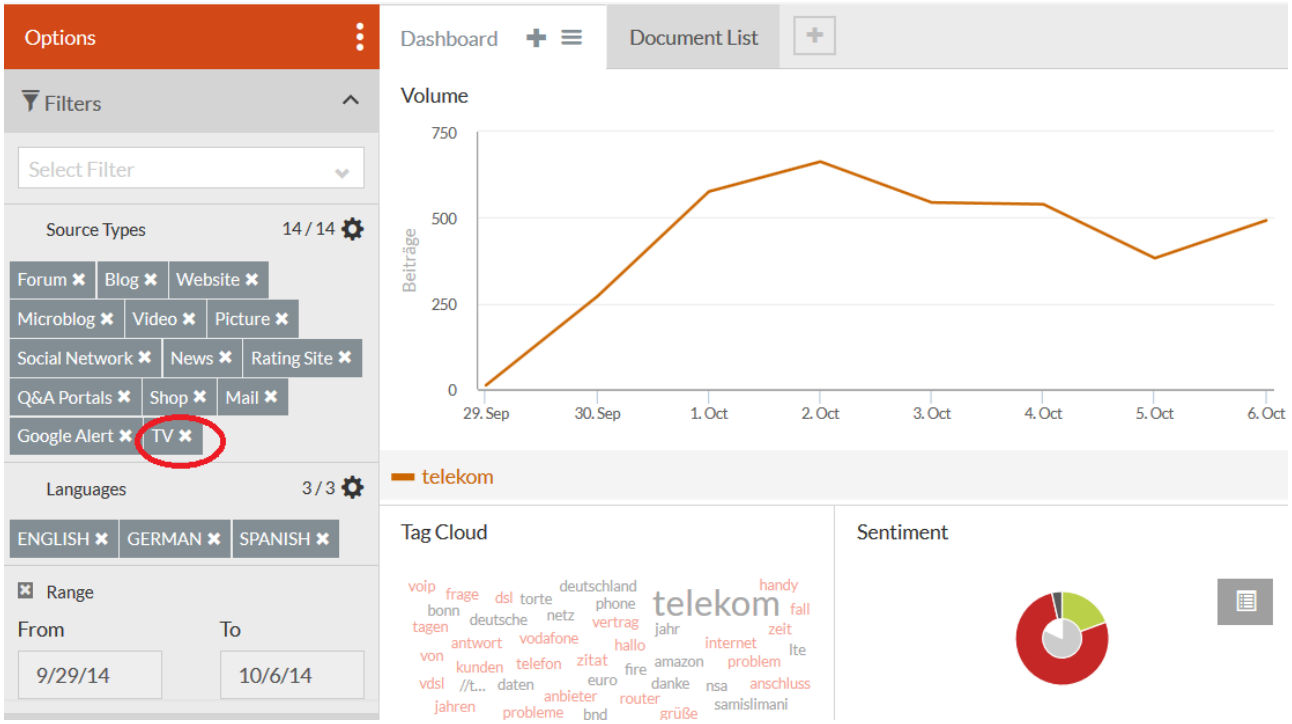

Figure 1: The VICO Media Monitoring Dashboard with the new data source type "TV"

.

Beneath the options bar in the dashboard (Figure 1) different options for the data selection can be chosen. These include Source Type (Social Network, TV, News, etc.), languages and feeds. Feeds contain queries to match a brand or a topic being monitored. In the example in Figure 1 one feed was chosen, containing the query *telekom*. It is also possible to select more feeds at the same time in order to compare results for different subjects of interest.

Figure 1 also shows a small selection of available analysis widgets in the dashboard. The *Chart* widget visualizes the data volume over time, showing the number of posts matching the selected feed(s) from selected sources and languages. A tag cloud can be generated from significant nouns as in Figure 1, but also from adjectives or Twitter hashtags. The system attempts to assign one of four sentiment values (i.e., positive, neutral, negative, mixed positive and negative) to all posts. Some assignments, however, might not be possible according to the underlying model and thus certain posts remain unassigned. The donut chart in the *Sentiment* widget in figure 1 describes the results of the sentiment analysis: the pie chart in the middle shows the portion of posts with assigned sentiment (white part) in relation to posts without a sentiment (grey part). The coloured chart shows the percentage of positive posts (green), negative posts (red), and mixed posts (dark grey) of all sentiment-assigned posts.

The dashboard offers a number of other widgets as well as many more filtering and analysis possibilities, which are beyond the scope of this document.

4 Use Case

To demonstrate the current version of the prototype we chose three companies to monitor. These are *Telekom Deutschland GmbH*, *Bayer AG,* and *BASF SE*. Additionally, we selected a topic, *ISIS,* currently widely discussed across media and languages. The search on the data is a simple keyword search, i.e. we search for **telekom, bayer, basf,** and **isis** respectively.

The data comes from

a) social media (Twitter, Facebook, blogs, forums), delivered by VICO

b) online news articles, delivered by JSI

c) speech-to-text transcripts of TV-shows, delivered by Zattoo / JSI

The VICO monitoring system normally includes selected news portals from other data providers as well. In the prototype, however, we limit the data sources to the sources available to all partners in the xLiMe project, meaning we also restrict the social media data to the xLiMe data pool (see also Section 2).

A thorough evaluation is not meant to be part of this document (see Section 5) but in the following we describe the first impression and some key findings immediately visible as we performed the first tests.

The covered languages are English, German, and Spanish. For the description in this document we look at data published between 2014-10-13 and 2014-10-20. This data set consists of almost 700,000 documents of which there are ~410,000 news articles, ~173,000 tweets, ~64,000 Facebook posts and comments, ~19,000 TV speech-to-text transcripts and ~25,000 documents from social media sources other than Twitter and Facebook.

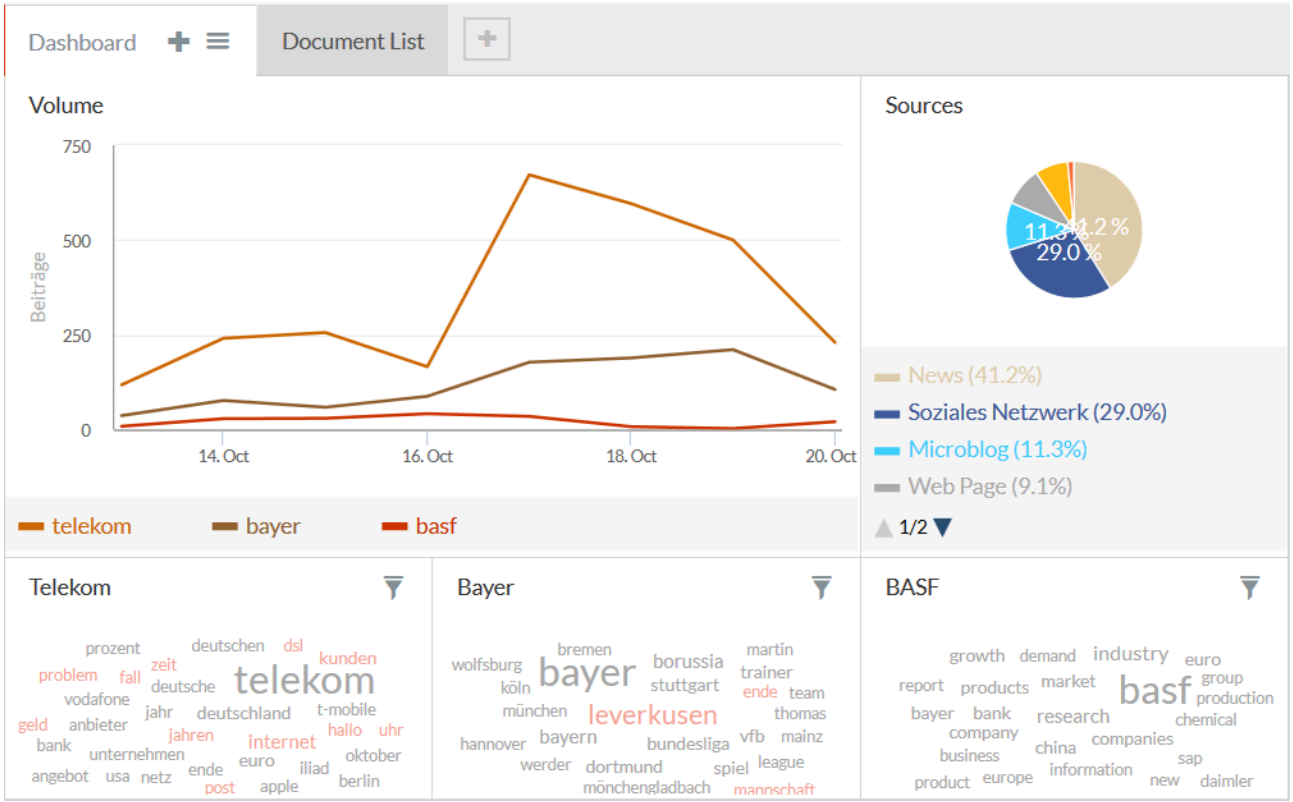

Figure 2: A part of a dashboard view containing a volume chart, a source chart, a tag cloud and a sentiment chart for the keywords bayer, basf, and telekom

First we take a look at a dashboard view showing results for the company names (Figure 2). The volume chart shows clearly that the *telekom* keyword has the most hits, with a peak on the 17th of October. In this particular dataset most documents represent a news article, followed by "Soziales Netzwerk" (social

network) and Microblog (Twitter). Only 0.1% of the documents are TV transcripts, this number is hidden in Figure 2.

In figure 2 there is a tag cloud for each of the keywords. The red words predominantly occur in posts with a negative sentiment assignment.

The next dashboard view (Figure 3) shows the results for the monitored concept *ISIS*. Here we have a slightly higher portion of TV-documents than for the company monitoring (i.e., 0.5% of all documents). There are more tweets than news articles for this subject, but the portion of social network posts is a lot smaller than for the company monitoring (12.8% vs. 29.0% for companies).

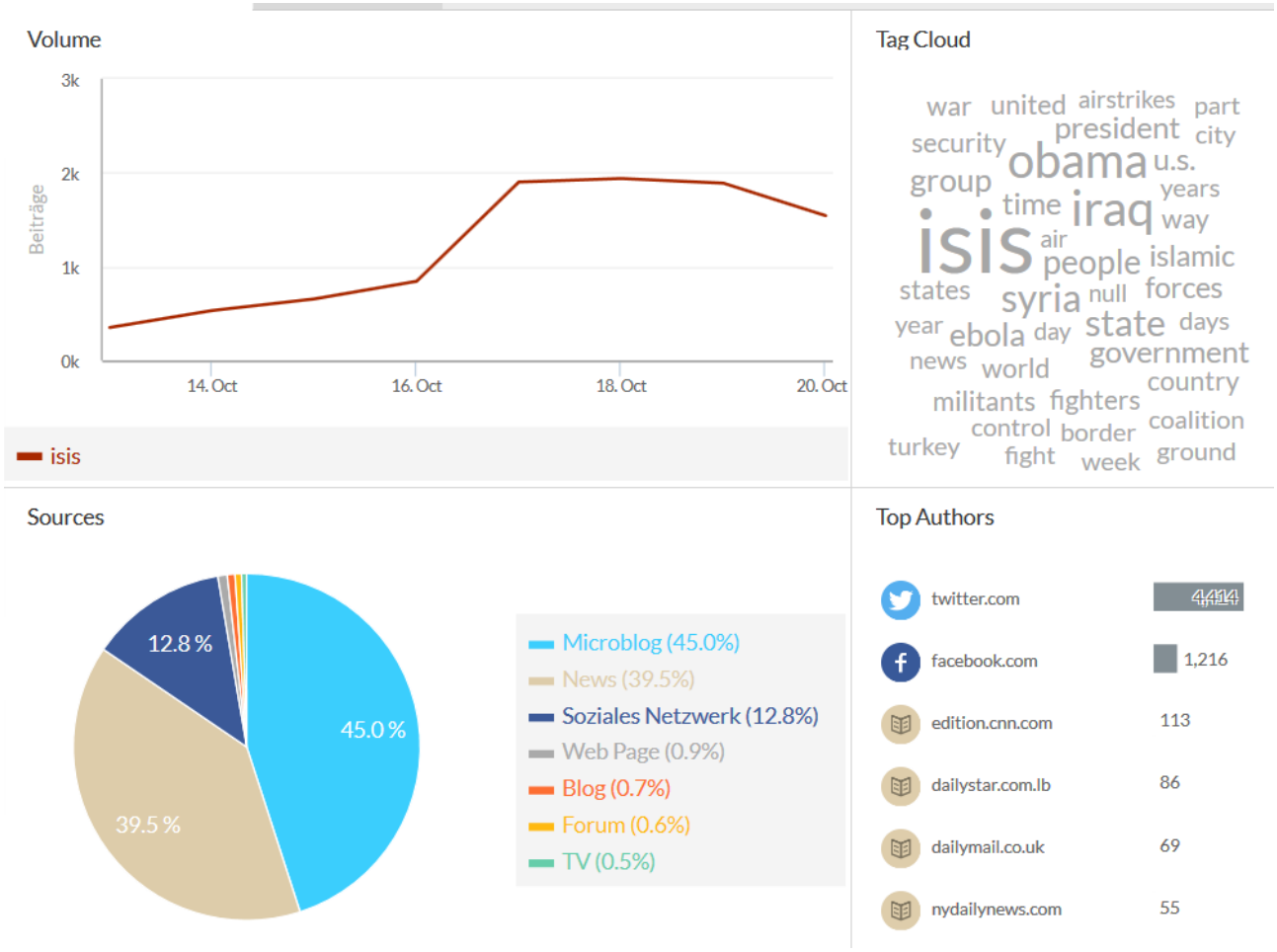

Figure 3: A part of a dashboard view containing a volume chart, a source chart, a tag cloud and a top authors list for the keyword isis

As we are particularly interested in the new source type TV, the final screen shot in Figure 4 shows a document view of TV speech-to-text transcripts. This view is also for the *ISIS* keyword which is highlighted in the document texts. Each transcript represents a 40s snippet from a TV-show. In Figure 4 there are two snippets from *The World Right Now with Hala Gorani* and one from *Connect the World with Becky Anderson.* Both shows run on CNN International.

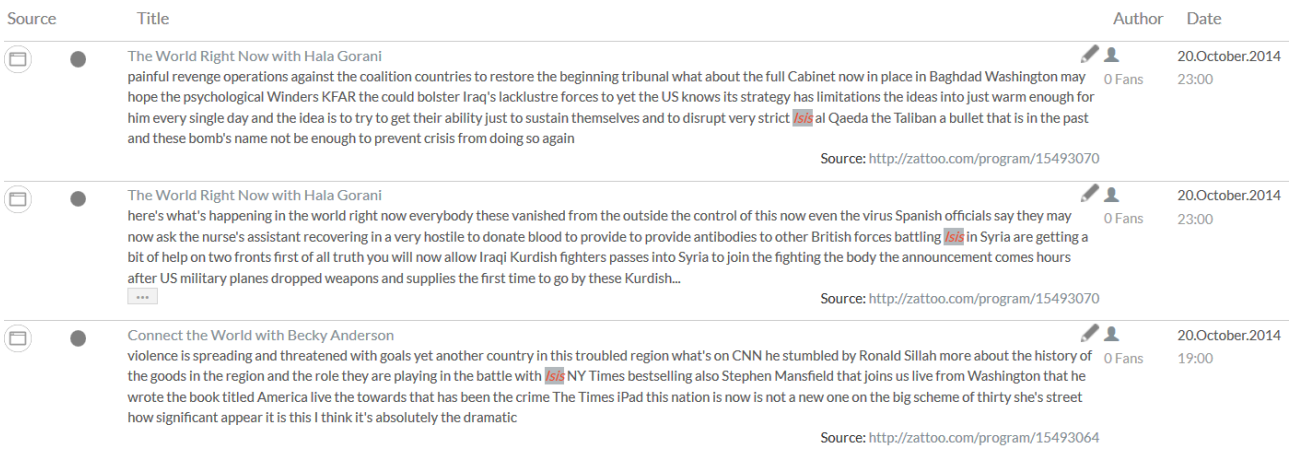

Figure 4: A part of a dashboard view containing a document list of TV speech-to-text transcripts matching the keyword isis

This short demonstration of integration of different sources of the xLiMe project shows that more data is needed from TV sources to make an impact on the overall analysis, even for a currently omnipresent subject like *ISIS*. Nevertheless, an individual examination of results from TV already has the potential to add to the insights of analysts and customers.

For a more comprehensive analysis and evaluation we will aim at gathering more data for more significant results.

5 Evaluation

Since the complete set of applied data sources has only been available for a short time, we left the formal evaluation process to M13 of the project. By then we will have a sufficient amount of data to be able to make more reliable statements on the year one early prototype. This document describes only the plans for the evaluation process, leaving the reporting of results to deliverable D1.4.2. The evaluation of the Y1 prototype will focus on measuring the impact the additional data sources (TV and news articles from JSI) have on the monitoring results. Hence, given a set of queries and a certain time frame, two questions are to be answered:

1. How much volume, in terms of the number of documents, do the additional data sources add to the results?
2. Do the results meet the needs of the users? That is, are the results related across media types and/or do the mainstream media sources add substantially to the information content of the resulting analysis? In particular, do they lead to any new insights on the subject of interest, not extractable from the results of social media data?

The evaluation will be performed by experienced analysts at VICO whereas external customer evaluation is planned at the latest for the Y2 prototype. Ideally, the customers will be involved already during the development of the Y2 prototype.

6 Conclusion

In this deliverable we described the early prototype version of the VICO use case "Monitor" of the xLiMe project. This version shows the first attempts to integrate cross-media sources into VICO's Social Media Monitoring System.

The xLiMe data processing pipeline was implemented, including providing of annotated social media data from VICO and the consuming and converting of data from the xLiMe data pool. The current monitoring system was extended to deal with a new media type and the first experiments have been made to include mainstream media into the analysing process. Further, the implementation of entity based search is in process. We described the planned evaluation procedure which will follow in M13 of the project.

References

[1] http://www.vico-research.com

[2] Early Speech To Text Prototype (Deliverable D2.1.1)

[3] http://www.ijs.si

[4] xLiMe Proposal (FP7-ICT-2013-10) Part B, p. 26

[5] xLiMe Proposal (FP7-ICT-2013-10) Part B, p. 10

[6] Prototype of Data Processing Infrastructure (Deliverable D1.2)

[7] Toolkit Architecture Specifications (Deliverable D6.1)

[8] http://corporate.zattoo.com

[9] Early Text from Video Prototype (Deliverable D2.2.1)

[10] Early Prototype for Video Annotations (Delvierable D3.2.1)

[11] Lei Zhang, Achim Rettinger, Michael Färber, Marko Tadić, A Comparative Evaluation of Cross-Lingual Text Annotation Techniques, “Information Access Evaluation. Multilinguality, Multimodality, and Visualization”, “Lecture Notes in Computer Science”, Volume 8138 2013, 124-135, 978-3-642-40801-4 (Print) 978-3-642-40802-1 (Online), 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain.

[12] Prototype of the (Meta) Data Model (Deliverable D1.1)

