The vision of the XLike project is to develop technologies to monitor and aggregate knowledge spreading across global mainstream and social media and to enable cross-lingual services for publishers, media monitoring and business intelligence. To achieve this we are combining scientific insights from several scientific areas to contribute in the area of text understanding.

The project is aiming to solve the following two open research problems:

1. **Extraction and integration of knowledge from multilingual texts with cross-lingual knowledge bases,** and
2. **Adapt linguistic techniques and crowdsourcing to deal with irregularities in informal language used primarily in social media.**

To solve the above two target research problems, the consortium brings expertise and insights into the project from several research areas including language technologies, cross-lingual technologies, machine learning, text mining, information retrieval, semantic and knowledge technologies.

The key tangible result of the project will be the “X-LIKE Software Toolkit”, which will serve as a basis for the use case applications. Functionally, the X-LIKE Toolkit consists from a pipeline of six separated stages:

- Multilingual Data Feeds
- Multilingual Linguistic Processing
- Cross-lingual Annotation
- Cross-lingual Semantic Integration
- Reporting and Analytics
- Toolkit and Interface

Figure 1. Schematic presentation of semantic Interlingua as an island bridged from the islands of textual information in various languages
XLiKe Major Achievements

Project produced several significant advancements towards the objectives in the second year of the project:

1. Project improved data infrastructure based around Newsfeed service by extending it to include 250,000 blogs, Twitter sprinkler feed, and Bloomberg and Slovenian Press Agency feeds. This required change to the crawling infrastructure in order to support large increase in new items. Newsfeed service was integrated with XLike pipeline. Now each article from one of the XLike languages is sent through multilingual linguistic processing, cross-lingual annotation and cross-lingual document linking services, and the Newsfeed contains all the enrichments produced by these services. This processed feed is the main input to the Event Registry and use-cases.

2. Project deployed new syntactic models for all XLike languages with improved performance. We have also deployed methods for semantic role labeling for English, Spanish, German, Catalan and Chinese. These are statistical models that recognize predicate-argument relations after syntactic analysis. These models constitute the backbone linguistic structure used in order to extract relational information that is relevant for the project use cases. In addition, and complementary to semantic role analysis, we have also developed rule-based extraction methods that allow to manually specify relation-extraction patterns. These will be useful for languages where we do not have resources to develop statistical semantic role systems, such as Slovene, as well as to extract specific patterns that may not follow predicative relations.

3. Project integrated existing tools for shallow analysis of micro-blogs in multilingual linguistic pipelines. We have implemented robust systems to find entities in micro-blogs and have investigated text-normalization techniques also for micro-blogs. Finally, there is research work in progress to adapt deep linguistic systems to out-of-domain data – in this case we are adopting distributional representations of lexical data, and redefining syntactic models to use such representations.

4. Project developed a new version of annotation tool which can annotate documents with the knowledge resources such as DBpedia and other Linked Open Data sources, with added cross-lingual groundings and annotation tool that can annotate documents with these resources. We have also developed initial prototype of ontology based WSD with support for knowledge resources used within XLike.

5. Project investigated applicability of statistical machine translation approaches for annotation of documents. The purpose of the early machine translation based semantic annotation prototype, is to investigate whether the SMT systems could be used to translate from natural language into a formal language such as Cyc Language. We have described the experiment using the Moses SMT system suite and presented the evaluation of results.

6. Project developed extension of the cross-lingual document linking prototype from the first year to cover 100 languages by improving the novel scalable approach based on hub languages developed in the first year.

7. Project developed two approaches for cross-lingual semantic graph construction from text. The first is based on linguistic infrastructure from WP2 and considers the grammatical relations appearing in the document, and generalizes such relations to obtain a cross-lingual representation of the
The relational content of the document. The second is based on pattern rules that link textual patterns into logical patterns. In this case, the logical patterns are the building blocks of the semantic graphs representing documents. A tool has been developed to explore a large collection of linguistically analyzed documents and manually construct patterns to map relational content into a cross-lingual semantic representation.

8. Project development an approach for cross-lingual event extraction. In order to identify events we developed a clustering algorithm that is able to group articles based different article features. Each identified cluster of articles initially represents a separate event. Once a cluster of articles is identified we extract key information about the event – title of the event, time when it happened, location and main entities involved in it. Since each cluster contains only articles in the same language we also need to identify and merge it with clusters in other languages that are describing the same event. We achieve this using the cross-lingual document linking approach. After the processing of an event is complete, we store it in the event registry.

9. Project developed the final visualization prototype by adding several new components: entity searching in the searching box, and improved time trends, languages and keywords charts. The enlarged time trend enables user to track articles at any timestamp in the selected time duration. The publisher bar shows a longer list of publishers and enables user to track articles of any selected publisher and so does the enlarged language pie.

10. The project developed a demonstrator, integrating all the technology developed within the first and second year. The prototype was successfully deployed and validated within Bloomberg and STA production environment. The validation results provided positive and valuable feedback.
Event Registry

**Event Registry** is a system that can analyse news articles and identify in them mentioned world events. For each event it can extract from the articles the main information about the event (who, when, where). Information about the events is stored in the Event Registry and can be accessed through a website.

The collected articles are pre-processed using XLike linguistic and annotation services in order to identify the entities and concepts mentioned in the articles. Articles with similar content in other languages are also identified using cross-lingual document linking.

The articles and the information extracted from them are then used to identify events described in the articles. An online-clustering method is used to group articles that are about the same story. Using the cross-lingual article matching and other features of the groups we can also connect identified events across languages and merge them into a single event. From articles in each event we then extract information about the event such as date, location and relevant entities.

Extracted information about the event is stored in the event registry. Event registry provides a user interface where users can search for events based on various criteria. Search results are visualized and aggregated on different ways in order to enable additional insights.

Content Advertising

Social media provides a novel channel for news publishing companies to directly engage with the users and drive new traffic to the website. In XLike we developed a prototype for Bloomberg use-case which can aid social media editors to optimizing content advertising through social media feeds. The prototype can identify articles for which there is demand in specific geographical regions (e.g. Germany, Italy) and offer them as suggestions to the editors, allowing them to target specific regions to increase engagement. The prototype monitors extrinsic signals (articles published by local news publishers and blogs in the local language) and intrinsic signals (real-time traffic monitoring on the Bloomberg.com website) to identify the demand for each article. Extrinsic measures are built on top of Newsfeed news crawler, and can link Bloomberg articles to content in local languages by using the cross-lingual document linking techniques developed in XLike. The result is a list of suggestions, which are available to social media editors through XLike HootSuite app.