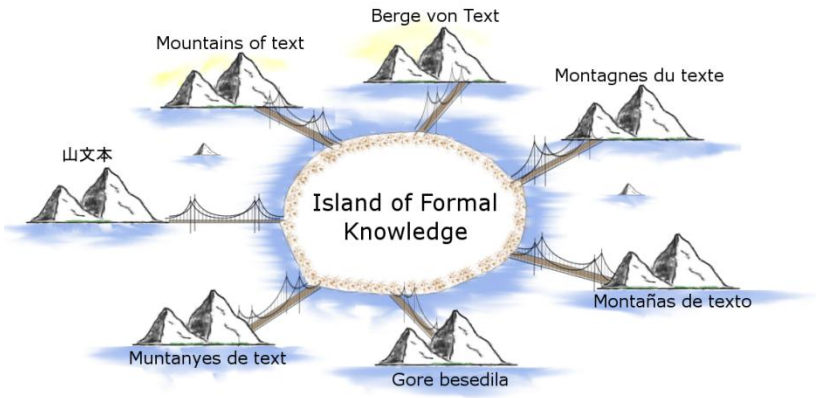
XLike

Cross-lingual Knowledge Extraction

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 aim is to combine scientific insights from several scientific areas to contribute in the area of cross-lingual text understanding. By combining modern computational linguistics, machine learning, text mining and semantic technologies we plan to deal with the following two key open research problems:

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

The developed technology will be language-agnostic, while within the project we will specifically address **English**, **German**, **Spanish**, **Chinese** and **Hindi** as major world languages and **Catalan** and **Slovenian** as minority languages.

SEVENTH FRAMEWORK PROGRAMME

Funded under: FP7

Area: Language Technologies (ICT-2011.4.2)

Project reference: 288342

Total cost: 4.57M euro

EU contribution: 3.55M euro

Duration: from January 2012 to December 2014 (36 months)

Contract type: Small and medium scale focused research project (STREP)

Coordinator: Marko Grobelnik

Website: www.xlike.org

Partners:

Jožef Stefan Institute
Karlsruhe Institute of Technology
Technical University of Catalonia
University of Zagreb
Tsinghua University
Intelligent Software Components
Bloomberg
Slovenian Press Agency

Associated Partners:

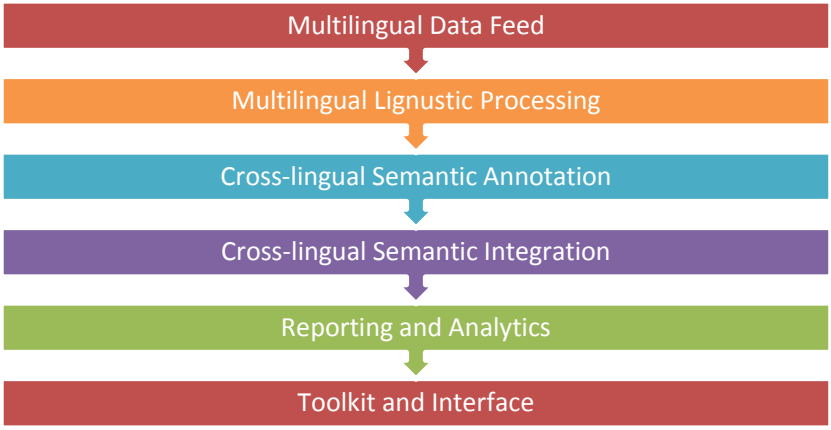
New York Times
IIT Bombay

Interlingua

Knowledge resources from Linked Open Data cloud will be used with special focus on general common sense knowledge base CycKB will be used as Interlingua. For the languages where no required linguistic resources will be available, we will use a probabilistic Interlingua representation trained from a comparable corpus drawn from the Wikipedia.

Project Pipeline

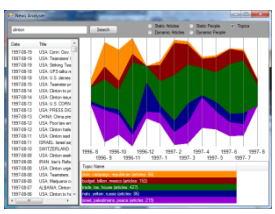
Schematic diagram of functional stages in the pipeline:

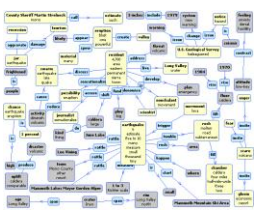


Use cases

The technology developed in the project will be used to develop solutions for cross-lingual summarization, contextualization, personalization and plagiarism detection of news stories with respect to global mainstream and social media articles.

Developed solutions will be applied and evaluated in two use cases: **Bloomberg** use case, covering the domain of financial news, and **Slovenian Press Agency**, covering the domain of general news.





Partners

Institut "Jožef Stefan" Ljubljana, Slovenija

Karlsruhe Institute of Technology

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAS STUDIORUM ZAGRABIENSIS MDCLXIX

TSINGHUA UNIVERSITY 1911

iSOCO enabling the networked economy

Bloomberg

sta

Associated Partners

The New York Times

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY