Deliverable D3.1.2

Final text annotation prototype

Editor:	Achim Rettinger, KIT
Author(s):	Lei Zhang, KIT; Achim Rettinger, KIT;
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)[1]	Public (PU)
Contractual Delivery Date:	M21
Actual Delivery Date:	1.10.2013
Suggested Readers:	All partners using the XLike Toolkit
Version:	0.1
Keywords:	Text annotation, Cross-lingual groundings, DBpedia, Wikipedia

[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.

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:	XLike – Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP3 – Cross-lingual Semantic Annotation
Document Title:	D3.1.2 Final text annotation prototype
Editor (Name, Affiliation)	Achim Rettinger, KIT
Work package Leader (Name, affiliation)	Achim Rettinger, KIT
Estimation of PM spent on the deliverable:	7 PM

Copyright notice

Executive Summary

The main goal of the XLike project is to extract knowledge from multilingual text documents by annotating textual unites, such as words and phrases, in the documents with a cross-lingual knowledge base. As the early text annotation prototype is to investigate the performance of shallow multilingual text annotation tools for knowledge base with pre-existing cross-lingual groundings. This deliverable provides the final version of 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. Within the task T3.1 we need to perform light weight approximate annotation between text and selection of relevant knowledge resources. Since this is the final deliverable for task T3.1, the results of this task, i.e., knowledge resources with cross-lingual lexical groundings and tool for text annotation, are both provided in this deliverable.

From now on whenever we use the term “XLike languages” we refer to English, German, Spanish, Chinese, Catalan and Slovenian.

Table of Contents

List of Figures

List of Tables

Abbreviations

STA	Slovenian Press Agency
BLP	Bloomberg
RDF	Resource Description Framework

1 Introduction

1.1 Motivation

The main goal of the XLike project is to extract knowledge from multi-lingual text documents by annotating textual unites, such as words and phrases, in the documents with a cross-lingual knowledge base. As the early text annotation prototype is to investigate the performance of shallow multilingual text annotation tools with Wikipedia. This deliverable provides the final version of 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. While this prototype annotates word phrases in the text documents and link them to DBpedia, Wikipedia and other knowledge resources, the semantic triple graphs merging prototype will extract subject-predicate-object triples and link them to a semantic knowledge representation. Such triples are essential to being able to apply logical constraints specified in a knowledge base or extract events based on event patterns. We consider the text annotation prototype as a prerequisite for the following semantic triple graphs merging and event extraction prototypes and the following use cases are related to this task.

1.2 Entity Tracking in Bloomberg Use Case

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. 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.

The entity-tracking task in the Bloomberg use case is to generate a more up-to-date list of relevant company news, preferably from their home markets for each company. The task can be roughly defined in two steps as (1) detect mentions of the entity (i.e. company), 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.

1.3 Topic and Entity Tracking in STA Use Case

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 and text annotation.

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.

1.4 Event Identification in STA Use Case

The goal is to develop methodology to identify the event mentioned in the article and to describe it with a set of properties (such as time of the event, involved entities, keywords, etc.). The developed algorithms will be able to assign each article to an event. The identified event will be either new (when this will be the first article describing it) or existing (when we have already seen other articles describing it). Events will be stored in an event registry that will provide querying and editing functionality.

An event is defined as a collection of semantic facts/assertions with the focus on actions. The (subject-predicate-object) assertions can be represented as a semantic graph. There are two main steps to detecting events: 1) extracting semantic facts from text (i.e. semantic graph construction) and 2) automatically deciding what set of facts constitutes an event. In order to extract new event patterns from the documents and identify events from the semantic graphs using event patterns, we need to first annotate the documents with knowledge resources, which relies on the word-sense-disambiguation of the annotations.

2 Techniques for Final Text Annotation

2.1 Background Knowledge Base

In this section, we will first introduce the main knowledge bases involved in this deliverable, namely DBpedia and Wikiepdia.

2.1.1 DBpedia

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to make sophisticated queries against Wikipedia, and to link other data sets on the Web to Wikipedia data. We hope this will make it easier for the amazing amount of information in Wikipedia to be used in new and interesting ways, and that it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.

The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases.

Localized versions of DBpedia in 111 languages are provided. All these versions together describe 20.8 million things, out of which 10.5 million overlap (are interlinked) with concepts from the English DBpedia. The full DBpedia data set features labels and abstracts for 10.3 million unique things in 111 different languages; 8.0 million links to images and 24.4 million HTML links to external web pages; 27.2 million data links into external RDF data sets, 55.8 million links to Wikipedia categories, and 8.2 million YAGO categories. The dataset consists of 1.89 billion pieces of information (RDF triples) out of which 400 million were extracted from the English edition of Wikipedia, 1.46 billion were extracted from other language editions, and about 27 million are data links into external RDF data sets.

2.1.2 Wikipedia

The online encyclopedia Wikipedia is a vast, constantly evolving tapestry of richly interlinked textual information. To a growing community of researchers and developers it is an ever-growing source of manually defined concepts and semantic relations. It constitutes an unparalleled and largely untapped resource for natural language processing, knowledge management, data mining, and other research areas.

All of Wikipedia's content is presented on pages of one kind or another, such as article, category, and redirect. Articles supply the bulk of Wikipedia's informative content. Each article describes a single concept or topic, and their titles are succinct, well-formed phrases that can be used as descriptors in ontologies and thesauri. Articles often contain links to equivalent articles in other language versions of Wikipedia. For example, the article Dog links to Chien in the French Wikipedia, Haushund in German, in Chinese, and many others.

Wikipedia provides several structural elements that associate articles with terms or surface forms that can be used to denote them. The most obvious elements are article titles and redirects: the article about dogs is given the title Dog and has about 30 redirects, including canis familiaris, domestic dog and man's best friend. The links made from other Wikipedia articles to this one provide additional surface forms, because authors tailor anchor text to suit the surrounding prose. A scientific article may contain a link to Dog that is labeled with the anchor text canis lupus familiaris, while a more informal article may refer to doggy.

Article titles, redirects and link anchors are all considered as Labels: terms (including phrases) that have been used to refer to the article in some way. Labels encode synonymy and polysemy, because it is possible for an article to be referred to by many labels, and for a label to refer to multiple articles. Because these labels are mined from an extensive corpus of text—that is, the full content of Wikipedia articles—they also have associated usage statistics: 76% of dog labels refer to the pet, 7% to the Chinese star sign, and less than 1% to hot dogs. How often labels are used within links, and how often they are found in plain text will be also tracked. The ratio between these statistics—prior link probability—helps to identify terms in new text that are likely to refer to concepts. For example, the probability that the term dog is linked in Wikipedia is 6%, while for the term the it is only 0.0006%. These statistics are useful for automatically detecting salient topics when they are mentioned in documents.

2.2 Cross-lingual groundings for DBpedia

We extract surface forms in all XLike languages of DBpedia resources from Wikipedia. Basically, we make use the following sources:

- Title of the page: Provide the most common name for resource
- Redirect page: Indicate synonyms, abbreviations or other variations of resource
- Disambiguation page: Useful in extracting abbreviations or other aliases of resource
- Anchor text: Very useful source of synonyms and other variations of resource

In order to derive cross-lingual grounds, we use cross-language links in Wikipedia, which connect "equivalent" resources across languages. Besides the extracted surface forms, we also exploit statistics of the cross-lingual groundings to answer the following questions:

- How important is a resource in different languages?
- How important is a surface form in different languages?
- How strong is a surface form associated with a resource?

In addition, we integrate all the lexical information described above into a DBpedia RDF triple store and allow users to query such information using SPARQL language. In the following we use some examples to introduce the RDF schema:

1. The count and probability of incoming page links of a DBpedia resource in Wikipedial

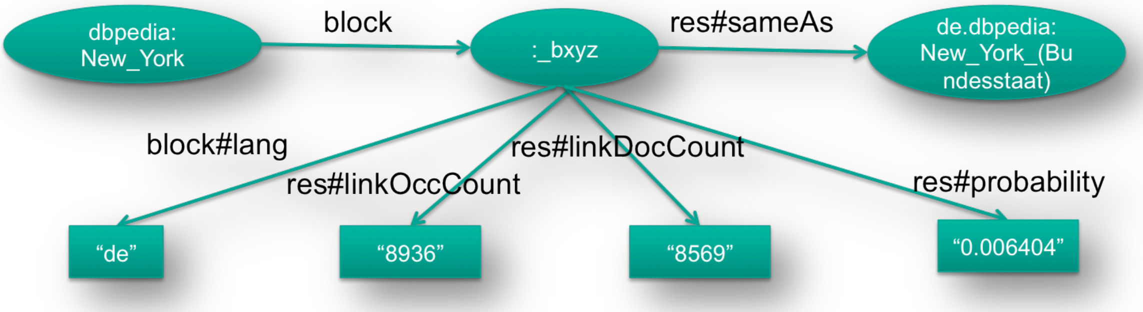


Figure 1. Resource block of the RDF schema

- <http://dbpedia.org/resource/New_York> <http://www.xlike.org/block> _:bxyz .
- _:bxyz <http://www.xlike.org/block#lang> "de" .
- _:bxyz <http://www.xlike.org/res#sameAs> <http://de.dbpedia.org/resource/New_York_(Bundesstaat)> .

The total number of links that are made to this resource in Wikipedia

- _:bxyz <http://www.xlike.org/res#linkDocCount> "8569"^^<http://www.w3.org/2001/XMLSchema#integer> .

The number of distinct articles, which contain a link to this resource in Wikipedia

- _:bxyz <http://www.xlike.org/res#linkOccCount> "8936"^^<http://www.w3.org/2001/XMLSchema#integer> .

The probability that this resource appears as a link in Wikipedia

- _:bxyz <http://www.xlike.org/res#probability> "0.006404"^^<http://www.w3.org/2001/XMLSchema#double> .

2. The count and probability of surface forms in different languages referring a DBpedia resource in Wikipedia

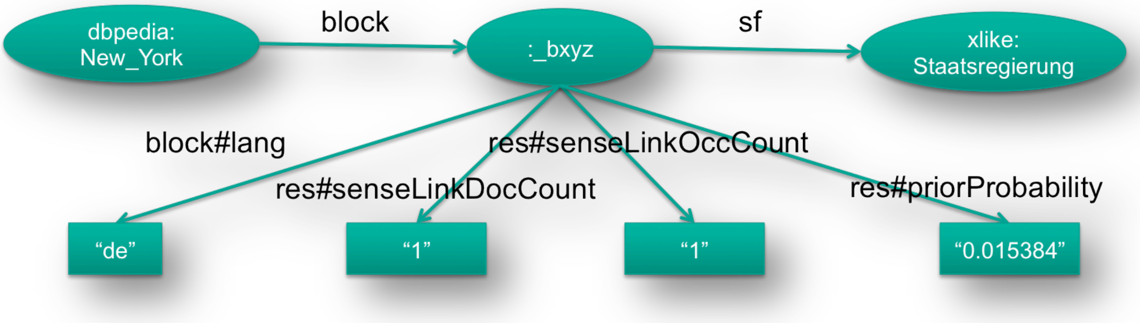


Figure 2. Sense block of the RDF schema

- <http://dbpedia.org/resource/New_York> <http://www.xlike.org/block> _:bxyz.
- _:bxyz <http://www.xlike.org/block#lang> "de" .
- _:bxyz <http://www.xlike.org/res#sf> <http://www.xlike.org/sf/Staatregierung> .

The number of documents that contain links that use the surrounding label as anchor text pointing to this sense as the destination

- _:bxyz <http://www.xlike.org/res#senseLinkDocCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .

The number of links that use the surrounding label as anchor text pointing to this sense as the destination

- _:bxyz <http://www.xlike.org/res#senseLinkOccCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .

The probability that the surrounding label goes to this destination

- _:bxyz <http://www.xlike.org/res#priorProbability> "0.015384"^^<http://www.w3.org/2001/XMLSchema#double> .

3. The statistics of a surface form

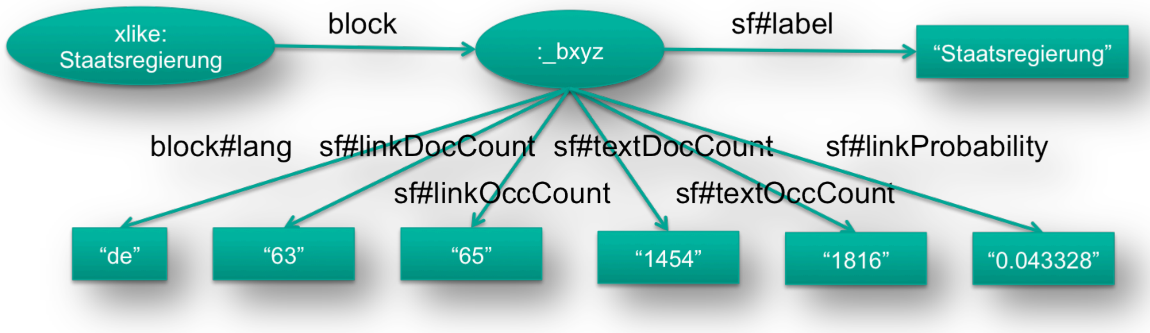


Figure 3. Surface form block of the RDF schema

- <http://www.xlike.org/sf/Staatsregierung> <http://www.xlike.org/block> _:bxyz .
- _:bxyz <http://www.xlike.org/block#lang> "de" .
- _:bxyz <http://www.xlike.org/sf#label> "Staatsregierung"@de .

The number of articles that contain links with this label used as an anchor

- _:bxyz <http://www.xlike.org/sf#linkDocCount> "63"^^<http://www.w3.org/2001/XMLSchema#integer> .

The number of links that use this label as an anchor

- _:bxyz <http://www.xlike.org/sf#linkOccCount> "65"^^<http://www.w3.org/2001/XMLSchema#integer> .

The number of articles that mention this label (either as links or in plain text)

- _:bxyz <http://www.xlike.org/sf#textDocCount> "1454"^^<http://www.w3.org/2001/XMLSchema#integer> .

The number of times this label is mentioned in articles (either as links or in plain text)

- _:bxyz <http://www.xlike.org/sf#textOccCount> "1816"^^<http://www.w3.org/2001/XMLSchema#integer> .

The probability that this surface form is used as a link in Wikipedia

- _:bxyz <http://www.xlike.org/sf#linkProbability> "0.043328"^^<http://www.w3.org/2001/XMLSchema#double> .

Multilingual information access can be facilitated by the availability of such cross-lingual lexicon, for example allowing for an easy mapping of natural language expressions in different languages to English ontology. We will introduce some usages here. First, given a DBpedia resource, we can retrieve its possible mentions in different languages together with the corresponding confidence score based on such lexicon. This will help for the following tasks:

- Natural language generation from RDF graph (RDF graph verbalization)
- Natural language generation from SPARQL queries (SPARQL query verbalization)

In addition, given a mention in any language, we can provide the resources that this mention might refer to and the corresponding confidence score. This will help for the following tasks:

- Cross-lingual entity linking
- Cross-lingual question answering

We can also assign a prior importance score to a mention. This will help for the following tasks:

- Criteria definition for entity linking benchmark

2.3 Final Text Annotation with Wikipedia and DBpedia

We first define the cross-lingual semantic annotation. On the one side, there are documents containing a set of entity mentions in source language. On the other side, there is a semantic knowledge base in target language containing a set of entities, each of which has its labels and attributes, and the relations between entities. Cross-lingual semantic annotation is to link/annotate these entity mentions in documents in the source language with their referent entities in the knowledge base in the target language. Here we assume that the cross-lingual groundings as discussed before are already integrated into the knowledge base.

In order to recognize mentions and disambiguate their meaning, generating DBpedia annotation in text, we will use 4 steps pipeline: 1) resource mention recognition, 2) candidate resource selection, 3) disambiguation graph construction, 4) graph-based disambiguation and 5) linking to other datasets.

In the first step, we gather word n-grams with maximal length from the document and match them with all entity surface forms to ascertain which terms and phrases correspond to entities in knowledge base. Each entity in the knowledge base is characterized by

- e.S: a collection of surface forms of an entity
- e.c: context of an entity

Each name mention m in a document is characterized by

- m.s: name of a mention
- m.c: context of a mention

In the second step, we extract a list of candidate entities with semantic similarity scores for each mention. The semantic similarity SS(m,e) between the mention m and the entity e can be computed as

$$SS(m,e) = \alpha \cdot LP(m,e) + \beta \cdot CS(m,e)$$

where LP(m,e) is the prior link probability of e for m and CS(m,e) is the context similarity between m and e.

In the third step, we add all detected mentions M = {m_1,...,m_p} and their candidates entities E = {e_1,...,e_q} as nodes into the disambiguation graph G. For each mention node m and one of its candidate entity node e, add an edge m -> e into G. For each pair of entity node e_i and e_j, add an edge e_i -> e_j into G, if there is an edge from e_i to e_j in knowledge base. The semantic relatedness between ei and ej is calculated based on Normalized Google Distance (NGD)

$$sr(e_i, e_j) = 1 - \frac{\log(\max(|E_i|, |E_j|)) - \log(|E_i \cap E_j|)}{\log(|N|) - \log(\min(|E_i|, |E_j|))}$$

where e_i and e_j are the two entities of interest, E_i and E_j are the sets of entities that link to e_i and e_j in the knowledge base respectively, and N is the set of all entities in the knowledge base.

After we obtain disambiguation graph G with mention nodes M = {m_1,...,m_p} and entity nodes E = {e_1,...,e_q}, and edges between them, we perform the personalized PageRank on the graph. The PageRank equation is shown as follows:

$$\mathrm{Pr} = \underbrace{c \cdot T \cdot \mathrm{Pr}}_{\text{Voting scheme}} + \underbrace{(1-c) \cdot \mathrm{v}}_{\text{Randomly jumping}}$$

$$T_{ij} = \begin{cases} \dfrac{sr(i,j)}{\sum_{k \in NB_i} sr(i,k)} & i,j \in E, i \rightarrow j \\ \dfrac{sim(i,j)}{\sum_{k \in E_i} sim(i,k)} & i \in M, j \in E, i \rightarrow j \\ 0 & otherwise \end{cases} \qquad v_i = \begin{cases} \dfrac{1}{p} & i \in M \\ 0 & otherwise \end{cases}$$

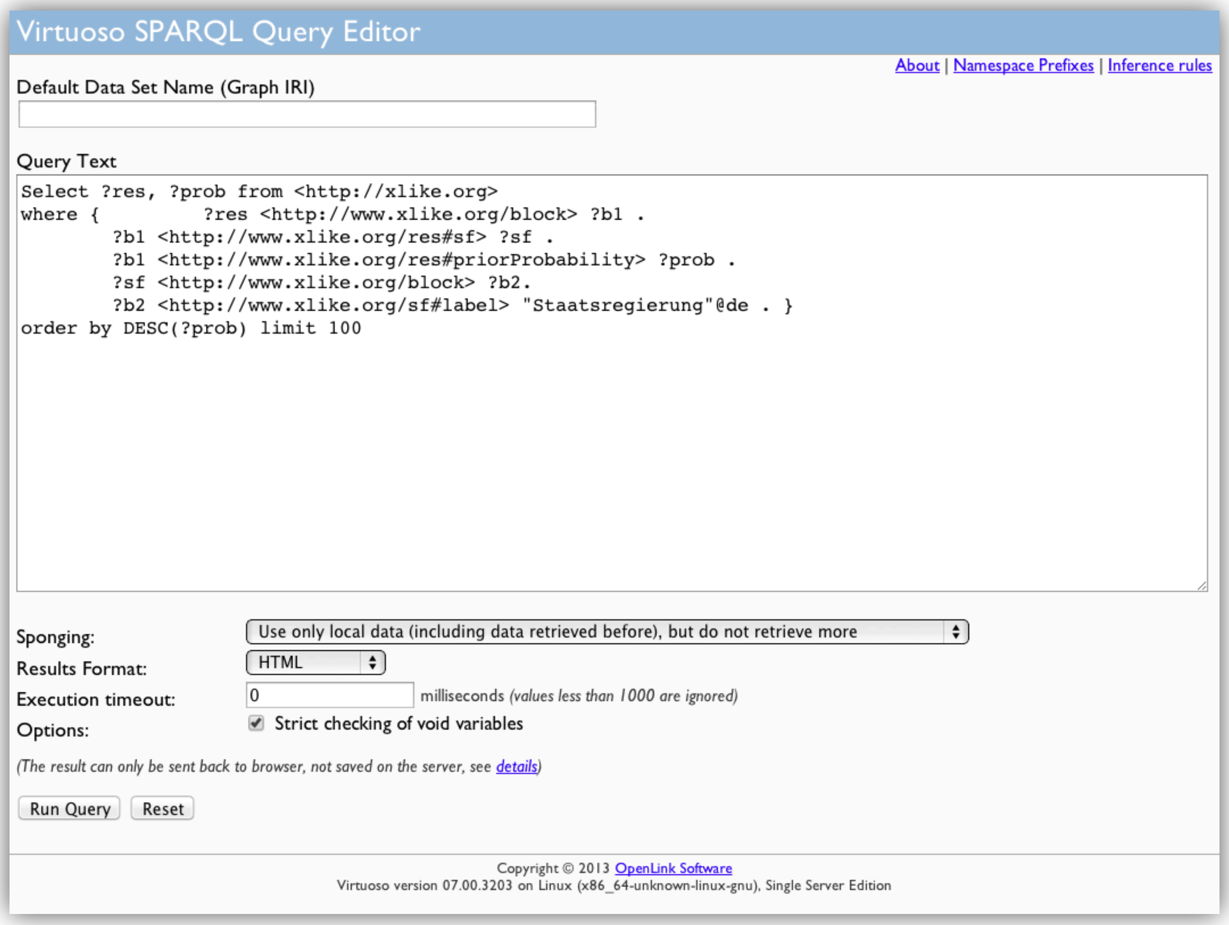
After performing Personalized PageRank on the disambiguation graph, we choose the entity with the highest probability for each mention. In the last step, we find the same resources in other datasets through *owl:sameAs* property contained in DBpedia.

3 Final Text Annotation Web Services

This section describes the technical implementation of the techniques introduced in the previous section.

3.1 DBpedia Cross-lingual Lexicon SPARQL Endpoint

Firstly, we introduce the SPARQL endpoint over the DBpedia lexicon data set, which provides cross-lingual lexical information about DBpedia resources. The endpoint is provided using OpenLink Virtuoso as the back-end database engine. This RDF data set used for this endpoint is extracted from Wikipedia dumps of July 2013 in English, German, Spanish, Catalan, Slovenian and Chinese. It contains 295 million triples of lexical information about DBpedia resources.


Virtuoso SPARQL Query Editor
About | Namespace Prefixes | Inference rules
Run Query
Reset

Default Data Set Name (Graph IRI)

Query Text

```
Select ?res, ?prob from <http://xlike.org>
where {          ?res <http://www.xlike.org/block> ?b1 .
        ?b1 <http://www.xlike.org/res#sf> ?sf .
        ?b1 <http://www.xlike.org/res#priorProbability> ?prob .
        ?sf <http://www.xlike.org/block> ?b2.
        ?b2 <http://www.xlike.org/sf#label> "Staatsregierung"@de . }
order by DESC(?prob) limit 100
```

Sponging: Use only local data (including data retrieved before), but do not retrieve more

Results Format: HTML

Execution timeout: 0 milliseconds *(values less than 1000 are ignored)*

Options: ☑ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

Copyright © 2013 OpenLink Software
Virtuoso version 07.00.3203 on Linux (x86_64-unknown-linux-gnu), Single Server Edition

Figure 4. Web User Interface of the DBpedia Cross-lingual Lexicon SPARQL Endpoint

The service and web user interface shown in Figure 4 are described in the following:

Service and web user interface URL: http://km.aifb.kit.edu/services/dbpedia-lexicon/

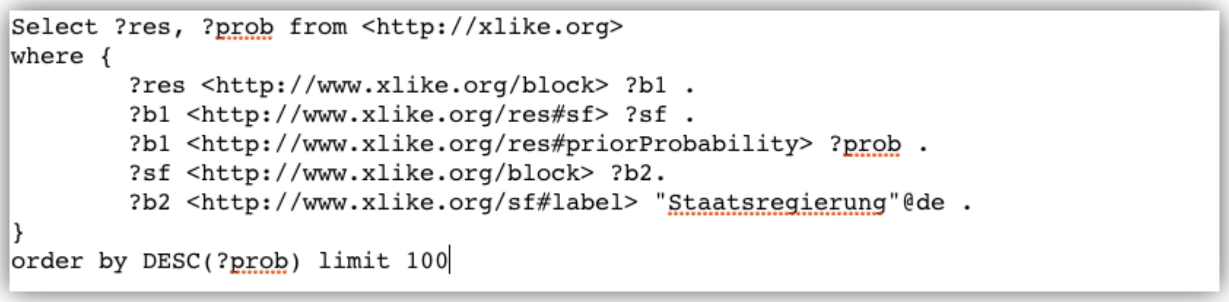
Input parameters:

- **query**: the SPARQL query
- **format**: xml, html, rdf, ntriples, json, csv ...

Output: the result of the SPARQL query in the format specified in the format paramete

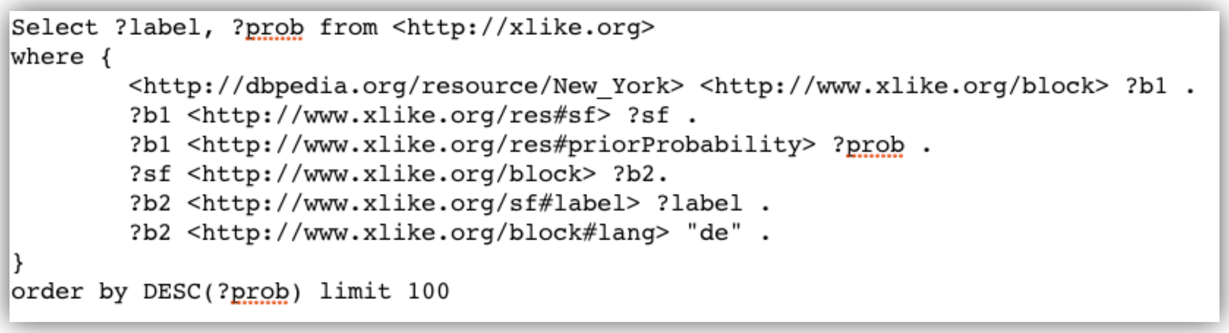
We list some examples of SPARQL queries to show how to use the endpoint:

Example 1 – Find top-k resources for a surface form in a language



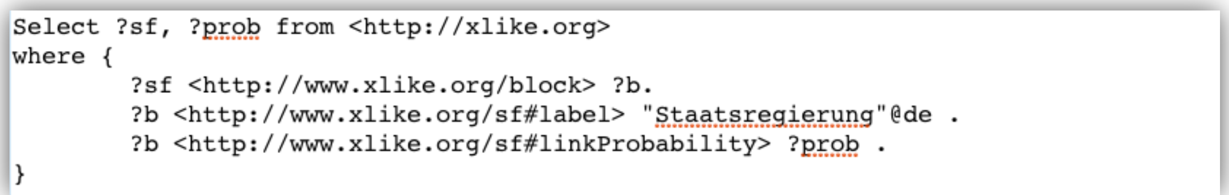
```
Select ?res, ?prob from <http://xlike.org>
where {
        ?res <http://www.xlike.org/block> ?b1 .
        ?b1 <http://www.xlike.org/res#sf> ?sf .
        ?b1 <http://www.xlike.org/res#priorProbability> ?prob .
        ?sf <http://www.xlike.org/block> ?b2.
        ?b2 <http://www.xlike.org/sf#label> "Staatsregierung"@de .
}
order by DESC(?prob) limit 100
```

Example 2 – Find top-k surface forms for a resource in a language



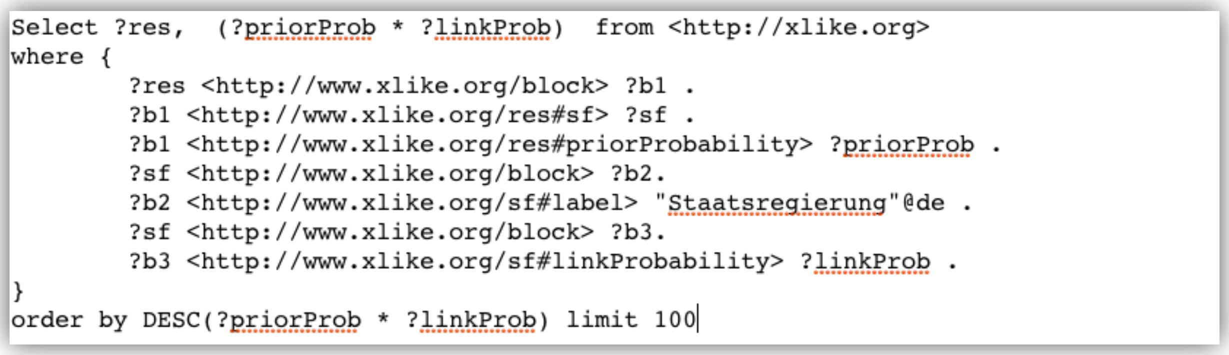
```
Select ?label, ?prob from <http://xlike.org>
where {
        <http://dbpedia.org/resource/New_York> <http://www.xlike.org/block> ?b1 .
        ?b1 <http://www.xlike.org/res#sf> ?sf .
        ?b1 <http://www.xlike.org/res#priorProbability> ?prob .
        ?sf <http://www.xlike.org/block> ?b2.
        ?b2 <http://www.xlike.org/sf#label> ?label .
        ?b2 <http://www.xlike.org/block#lang> "de" .
}
order by DESC(?prob) limit 100
```

Example 3 – Find the link probability of a surface form in a language



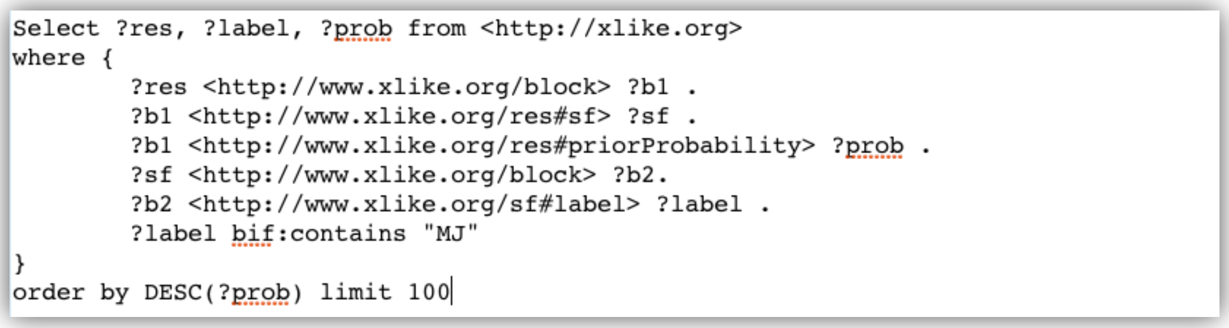
```
Select ?sf, ?prob from <http://xlike.org>
where {
        ?sf <http://www.xlike.org/block> ?b.
        ?b <http://www.xlike.org/sf#label> "Staatsregierung"@de .
        ?b <http://www.xlike.org/sf#linkProbability> ?prob .
}
```

Example 4 – Find top-k resources for a surface form using aggregate function



```
Select ?res,  (?priorProb * ?linkProb)  from <http://xlike.org>
where {
        ?res <http://www.xlike.org/block> ?b1 .
        ?b1 <http://www.xlike.org/res#sf> ?sf .
        ?b1 <http://www.xlike.org/res#priorProbability> ?priorProb .
        ?sf <http://www.xlike.org/block> ?b2.
        ?b2 <http://www.xlike.org/sf#label> "Staatsregierung"@de .
        ?sf <http://www.xlike.org/block> ?b3.
        ?b3 <http://www.xlike.org/sf#linkProbability> ?linkProb .
}
order by DESC(?priorProb * ?linkProb) limit 100
```

Example 5 – Find top-k resources with surface form containing “MJ”



```
Select ?res, ?label, ?prob from <http://xlike.org>
where {
        ?res <http://www.xlike.org/block> ?b1 .
        ?b1 <http://www.xlike.org/res#sf> ?sf .
        ?b1 <http://www.xlike.org/res#priorProbability> ?prob .
        ?sf <http://www.xlike.org/block> ?b2.
        ?b2 <http://www.xlike.org/sf#label> ?label .
        ?label bif:contains "MJ"
}
order by DESC(?prob) limit 100
```

3.2 Final Text Annotation Service in the XLike Pipeline

This web service takes the output of linguistic processing in WP2 as input, adds the annotations with DBpedia and Wikipedia resources on top of the input.

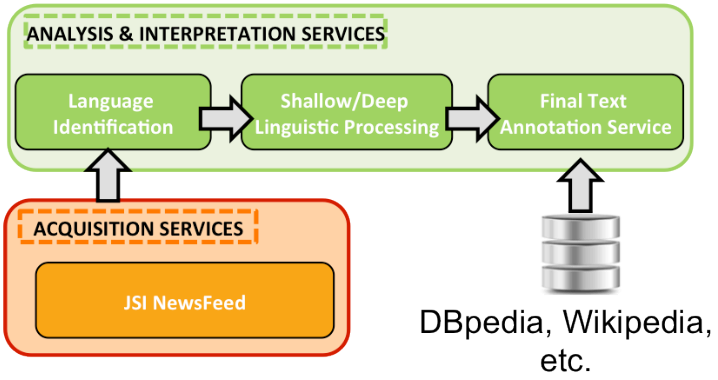
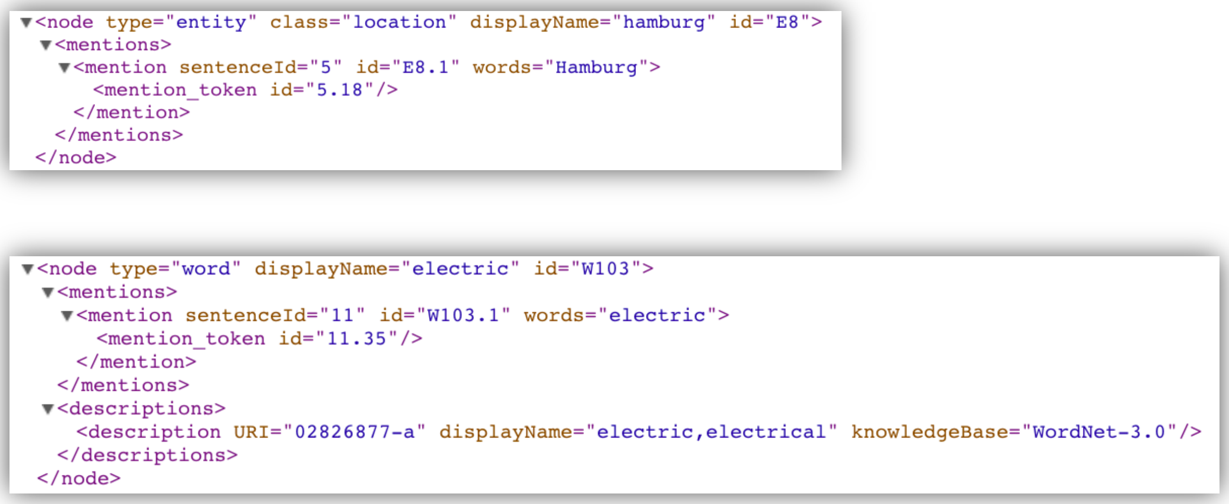
Service Name	**Final Text Annotation Service**
Description	This web service takes the output of multi-linguistic processing in WP2 as input and adds the annotations with knowledge resources, such as DBpedia and Wikipedia etc. by using the graph-based disambiguation algorithm, such as personalized PageRank.
URI	http://km.aifb.kit.edu/services/text-annotation-xx/ (where xx is the language)
Source Code Repository	Will be published to private XLike GitHub repository.
APIs Implemented	Parameters: Output of multi-linguistic processing in WP2
Services Used	 ANALYSIS & INTERPRETATION SERVICES Language Identification Shallow/Deep Linguistic Processing Final Text Annotation Service ACQUISITION SERVICES JSI NewsFeed DBpedia, Wikipedia, etc. The final text annotation service annotates the entities provided by the multilingual processing analysis with the knowledge resources for the different languages.
Additional Information	None
Notes	None

Table 1. Component of Final Text Annotation Service for DBpedia and Wikipedia

This web service takes the output of linguistic processing in WP2 as input, adds the DBpedia and Wikipedia annotations by using the graph-based disambiguation algorithm, such as personalized PageRank.

Language	**URL SandBox**	**Parameters**
English Service	http://km.aifb.kit.edu/services/text-annotation-en/	<item> <sentences> <sentence id=""> <text> </text> <tokens> <token pos=" " end="" lemma=" " id="" start=""> </token> </tokens> </sentence> </sentences> <nodes> <node type=" " displayName=" " id=""> <mentions> <mention sentenceId="" id="" words=" "></mention> <mention_token></mention_token> </mentions> </node> </nodes> <frames>...</frames> <conll>...</conll> </item>
Spanish Service	http://km.aifb.kit.edu/services/text-annotation-es/	Same as English Service
Catalan Service	http://km.aifb.kit.edu/services/text-annotation-ca/	Same as English Service
German Service	http://km.aifb.kit.edu/services/text-annotation-de/	Same as English Service
Chinese Service	http://km.aifb.kit.edu/services/text-annotation-zh/	Same as English Service
Slovenian Service	http://km.aifb.kit.edu/services/text-annotation-sl/	Same as English Service

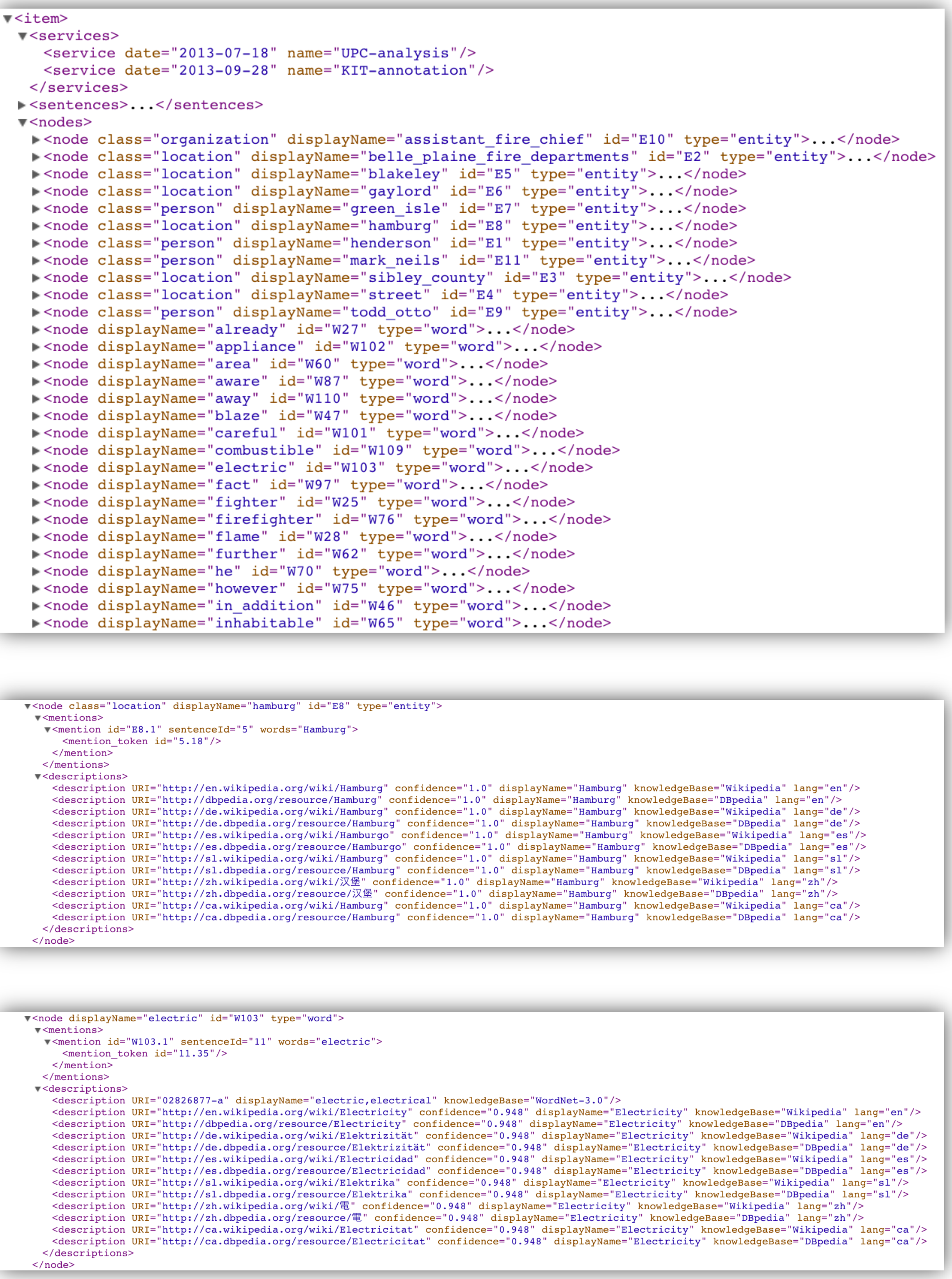
Table 2. Description of Final Text Annotation Service in the XLike pipeline



```
▼<node type="entity" class="location" displayName="hamburg" id="E8">
  ▼<mentions>
    ▼<mention sentenceId="5" id="E8.1" words="Hamburg">
       <mention_token id="5.18"/>
     </mention>
   </mentions>
 </node>
```

```
▼<node type="word" displayName="electric" id="W103">
  ▼<mentions>
    ▼<mention sentenceId="11" id="W103.1" words="electric">
       <mention_token id="11.35"/>
     </mention>
   </mentions>
  ▼<descriptions>
     <description URI="02826877-a" displayName="electric,electrical" knowledgeBase="WordNet-3.0"/>
   </descriptions>
 </node>
```

Figure 5. Example input of the text annotation service in the XLike pipeline



```
▼<item>
  ▼<services>
     <service date="2013-07-18" name="UPC-analysis"/>
     <service date="2013-09-28" name="KIT-annotation"/>
   </services>
  ▶<sentences>...</sentences>
  ▼<nodes>
    ▶<node class="organization" displayName="assistant_fire_chief" id="E10" type="entity">...</node>
    ▶<node class="location" displayName="belle_plaine_fire_departments" id="E2" type="entity">...</node>
    ▶<node class="location" displayName="blakeley" id="E5" type="entity">...</node>
    ▶<node class="location" displayName="gaylord" id="E6" type="entity">...</node>
    ▶<node class="person" displayName="green_isle" id="E7" type="entity">...</node>
    ▶<node class="location" displayName="hamburg" id="E8" type="entity">...</node>
    ▶<node class="person" displayName="henderson" id="E1" type="entity">...</node>
    ▶<node class="person" displayName="mark_neils" id="E11" type="entity">...</node>
    ▶<node class="location" displayName="sibley_county" id="E3" type="entity">...</node>
    ▶<node class="location" displayName="street" id="E4" type="entity">...</node>
    ▶<node class="person" displayName="todd_otto" id="E9" type="entity">...</node>
    ▶<node displayName="already" id="W27" type="word">...</node>
    ▶<node displayName="appliance" id="W102" type="word">...</node>
    ▶<node displayName="area" id="W60" type="word">...</node>
    ▶<node displayName="aware" id="W87" type="word">...</node>
    ▶<node displayName="away" id="W110" type="word">...</node>
    ▶<node displayName="blaze" id="W47" type="word">...</node>
    ▶<node displayName="careful" id="W101" type="word">...</node>
    ▶<node displayName="combustible" id="W109" type="word">...</node>
    ▶<node displayName="electric" id="W103" type="word">...</node>
    ▶<node displayName="fact" id="W97" type="word">...</node>
    ▶<node displayName="fighter" id="W25" type="word">...</node>
    ▶<node displayName="firefighter" id="W76" type="word">...</node>
    ▶<node displayName="flame" id="W28" type="word">...</node>
    ▶<node displayName="further" id="W62" type="word">...</node>
    ▶<node displayName="he" id="W70" type="word">...</node>
    ▶<node displayName="however" id="W75" type="word">...</node>
    ▶<node displayName="in_addition" id="W46" type="word">...</node>
    ▶<node displayName="inhabitable" id="W65" type="word">...</node>
```

```
▼<node class="location" displayName="hamburg" id="E8" type="entity">
  ▼<mentions>
    ▼<mention id="E8.1" sentenceId="5" words="Hamburg">
       <mention_token id="5.18"/>
     </mention>
   </mentions>
  ▼<descriptions>
     <description URI="http://en.wikipedia.org/wiki/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="en"/>
     <description URI="http://dbpedia.org/resource/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="en"/>
     <description URI="http://de.wikipedia.org/wiki/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="de"/>
     <description URI="http://de.dbpedia.org/resource/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="de"/>
     <description URI="http://es.wikipedia.org/wiki/Hamburgo" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="es"/>
     <description URI="http://es.dbpedia.org/resource/Hamburgo" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="es"/>
     <description URI="http://sl.wikipedia.org/wiki/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="sl"/>
     <description URI="http://sl.dbpedia.org/resource/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="sl"/>
     <description URI="http://zh.wikipedia.org/wiki/汉堡" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="zh"/>
     <description URI="http://zh.dbpedia.org/resource/汉堡" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="zh"/>
     <description URI="http://ca.wikipedia.org/wiki/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="Wikipedia" lang="ca"/>
     <description URI="http://ca.dbpedia.org/resource/Hamburg" confidence="1.0" displayName="Hamburg" knowledgeBase="DBpedia" lang="ca"/>
   </descriptions>
 </node>
```

```
▼<node displayName="electric" id="W103" type="word">
  ▼<mentions>
    ▼<mention id="W103.1" sentenceId="11" words="electric">
       <mention_token id="11.35"/>
     </mention>
   </mentions>
  ▼<descriptions>
     <description URI="02826877-a" displayName="electric,electrical" knowledgeBase="WordNet-3.0"/>
     <description URI="http://en.wikipedia.org/wiki/Electricity" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="en"/>
     <description URI="http://dbpedia.org/resource/Electricity" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="en"/>
     <description URI="http://de.wikipedia.org/wiki/Elektrizität" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="de"/>
     <description URI="http://de.dbpedia.org/resource/Elektrizität" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="de"/>
     <description URI="http://es.wikipedia.org/wiki/Electricidad" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="es"/>
     <description URI="http://es.dbpedia.org/resource/Electricidad" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="es"/>
     <description URI="http://sl.wikipedia.org/wiki/Elektrika" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="sl"/>
     <description URI="http://sl.dbpedia.org/resource/Elektrika" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="sl"/>
     <description URI="http://zh.wikipedia.org/wiki/電" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="zh"/>
     <description URI="http://zh.dbpedia.org/resource/電" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="zh"/>
     <description URI="http://ca.wikipedia.org/wiki/Electricitat" confidence="0.948" displayName="Electricity" knowledgeBase="Wikipedia" lang="ca"/>
     <description URI="http://ca.dbpedia.org/resource/Electricitat" confidence="0.948" displayName="Electricity" knowledgeBase="DBpedia" lang="ca"/>
   </descriptions>
 </node>
```

Figure 6. Example output of the text annotation service in the XLike pipeline

3.3 Service for Annotating Web Pages and Raw Texts

This web service is based on Explicit Semantic Analysis (ESA) and language links in Wikipedia. It uses Wikipedia dumps from May 2012 in English, German, Spanish, French, Catalan and Slovenian. The service can be called by using POST or GET request to the following URL address and input parameters.

Service URL: http://km.aifb.kit.edu/services/webpage-annotation/

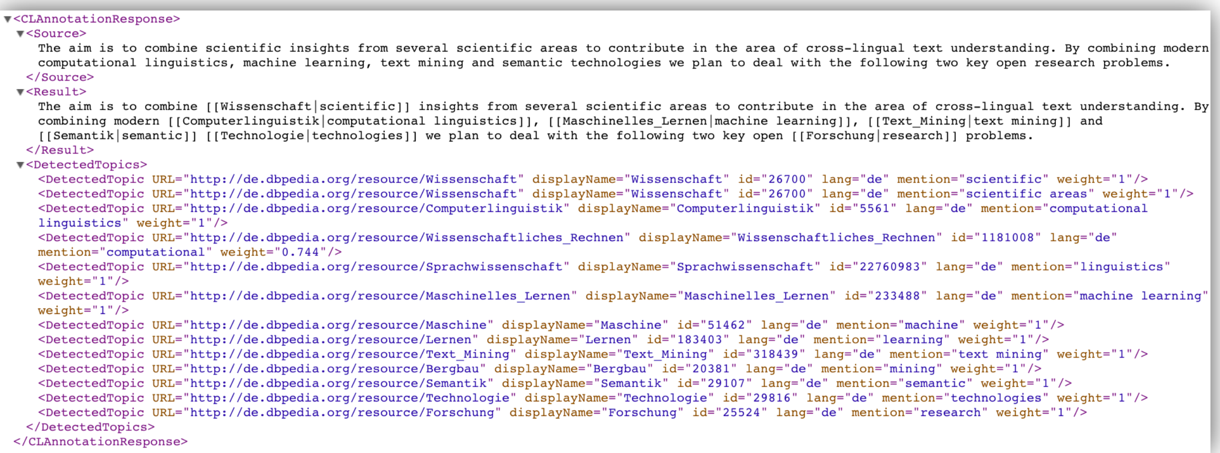
Input Parameters:

- **source**: the URL of a web page or raw text
- **lang1**: language of source information
- **lang2**: language of knowledge base resources
- **kb**: the knowledge base used for annotation (DBpedia or Wikipedia)

When the *source* parameter is raw text, the response of the service consists of the xml elements **`CLAnnotationResponse`**, which consists of a **`source`** element containing the raw input document before pre-processing, a **`result`** element containing annotated text and a **`DetectedTopics`** element, which consists of a list of the **`DetectedTopic`** elements. The **`DetectedTopicc`** element has the following attributes:

- **URL**: the URL of the knowledge base resource
- **id**: the id of the knowledge base resource
- **mention**: the mention of the knowledge base resource in the text
- **lang**: the language of the knowledge base resource
- **displayName**: the display name of the knowledge base resource
- **weigth**: measure of the confidence between the mention and the knowledge base resource

Figure 7 shows the output format of the service when the *source* parameter is raw text.

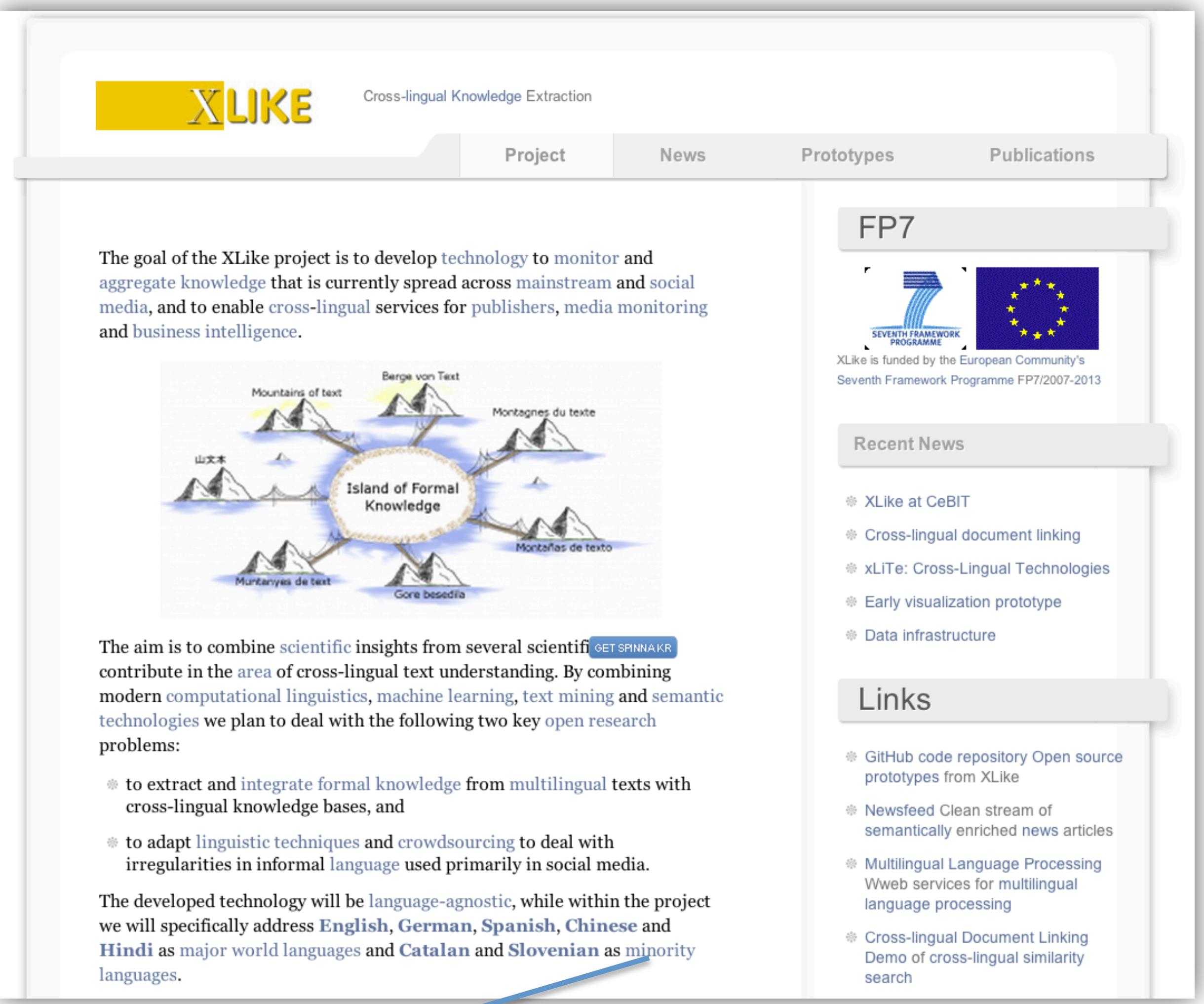


```
▼<CLAnnotationResponse>
  ▼<Source>
     The aim is to combine scientific insights from several scientific areas to contribute in the area of cross-lingual text understanding. By combining modern
     computational linguistics, machine learning, text mining and semantic technologies we plan to deal with the following two key open research problems.
   </Source>
  ▼<Result>
     The aim is to combine [[Wissenschaft|scientific]] insights from several scientific areas to contribute in the area of cross-lingual text understanding. By
     combining modern [[Computerlinguistik|computational linguistics]], [[Maschinelles_Lernen|machine learning]], [[Text_Mining|text mining]] and
     [[Semantik|semantic]] [[Technologie|technologies]] we plan to deal with the following two key open [[Forschung|research]] problems.
   </Result>
  ▼<DetectedTopics>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Wissenschaft" displayName="Wissenschaft" id="26700" lang="de" mention="scientific" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Wissenschaft" displayName="Wissenschaft" id="26700" lang="de" mention="scientific areas" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Computerlinguistik" displayName="Computerlinguistik" id="5561" lang="de" mention="computational
     linguistics" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Wissenschaftliches_Rechnen" displayName="Wissenschaftliches_Rechnen" id="1181008" lang="de"
     mention="computational" weight="0.744"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Sprachwissenschaft" displayName="Sprachwissenschaft" id="22760983" lang="de" mention="linguistics"
     weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Maschinelles_Lernen" displayName="Maschinelles_Lernen" id="233488" lang="de" mention="machine learning"
     weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Maschine" displayName="Maschine" id="51462" lang="de" mention="machine" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Lernen" displayName="Lernen" id="183403" lang="de" mention="learning" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Text_Mining" displayName="Text_Mining" id="318439" lang="de" mention="text mining" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Bergbau" displayName="Bergbau" id="20381" lang="de" mention="mining" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Semantik" displayName="Semantik" id="29107" lang="de" mention="semantic" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Technologie" displayName="Technologie" id="29816" lang="de" mention="technologies" weight="1"/>
     <DetectedTopic URL="http://de.dbpedia.org/resource/Forschung" displayName="Forschung" id="25524" lang="de" mention="research" weight="1"/>
   </DetectedTopics>
 </CLAnnotationResponse>
```

Figure 7. Example output of the service for annotating raw text

When the *source* parameter is URL of a web page, the response of the service is the web page with inserted annotation based on the resources in the knowledge base specified in the *kb* parameter.

Figure 8 shows the output format of the service when the *source* parameter is URL of a web page.



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.

The aim is to combine scientific insights from several scientifi 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.

FP7

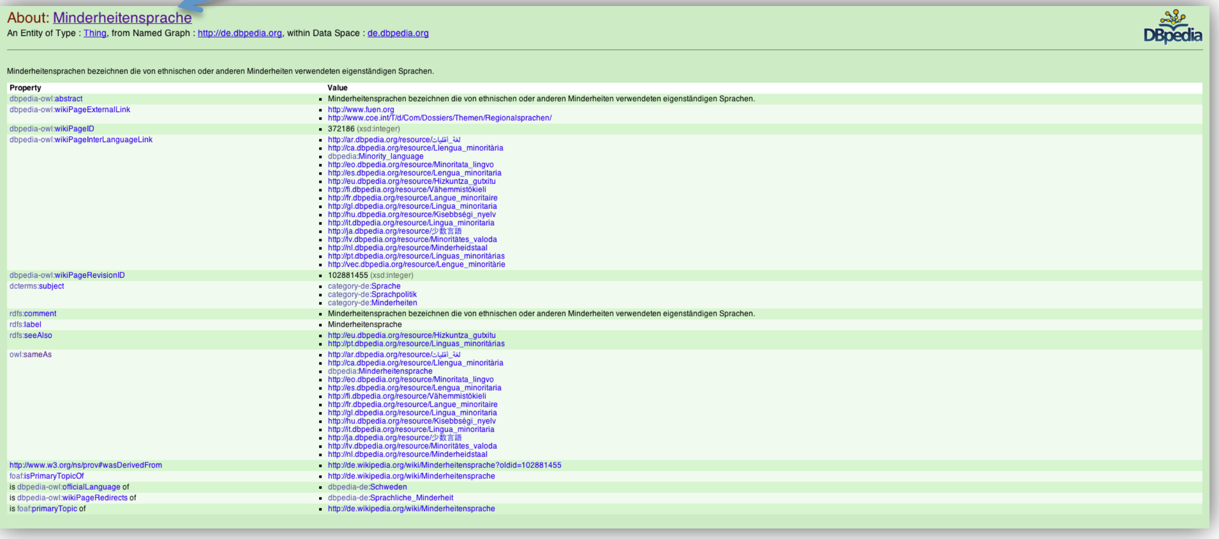
XLike is funded by the European Community's Seventh Framework Programme FP7/2007-2013

Recent News

- XLike at CeBIT
- Cross-lingual document linking
- xLiTe: Cross-Lingual Technologies
- Early visualization prototype
- Data infrastructure

Links

- GitHub code repository Open source prototypes from XLike
- Newsfeed Clean stream of semantically enriched news articles
- Multilingual Language Processing Wweb services for multilingual language processing
- Cross-lingual Document Linking Demo of cross-lingual similarity search



About: Minderheitensprache

An Entity of Type : Thing, from Named Graph : http://de.dbpedia.org, within Data Space : de.dbpedia.org

Minderheitensprachen bezeichnen die von ethnischen oder anderen Minderheiten verwendeten eigenständigen Sprachen.

Property	Value
dbpedia-owl:abstract	▪ Minderheitensprachen bezeichnen die von ethnischen oder anderen Minderheiten verwendeten eigenständigen Sprachen.
dbpedia-owl:wikiPageExternalLink	▪ http://www.fuen.org ▪ http://www.coe.int/T/d/Com/Dossiers/Themen/Regionalsprachen/
dbpedia-owl:wikiPageID	▪ 372186 (xsd:integer)
dbpedia-owl:wikiPageInterLanguageLink	▪ http://ar.dbpedia.org/resource/لغة_أقليات ▪ http://ca.dbpedia.org/resource/Llengua_minoritària ▪ dbpedia:Minority_language ▪ http://eo.dbpedia.org/resource/Minoritata_lingvo ▪ http://es.dbpedia.org/resource/Lengua_minoritaria ▪ http://eu.dbpedia.org/resource/Hizkuntza_gutxitu ▪ http://fi.dbpedia.org/resource/Vähemmistökieli ▪ http://fr.dbpedia.org/resource/Langue_minoritaire ▪ http://gl.dbpedia.org/resource/Lingua_minoritaria ▪ http://hu.dbpedia.org/resource/Kisebbségi_nyelv ▪ http://it.dbpedia.org/resource/Lingua_minoritaria ▪ http://ja.dbpedia.org/resource/少数言語 ▪ http://lv.dbpedia.org/resource/Minoritātes_valoda ▪ http://nl.dbpedia.org/resource/Minderheidstaal ▪ http://pt.dbpedia.org/resource/Línguas_minoritárias ▪ http://vec.dbpedia.org/resource/Lengue_minoritàrie
dbpedia-owl:wikiPageRevisionID	▪ 102881455 (xsd:integer)
dcterms:subject	▪ category-de:Sprache ▪ category-de:Sprachpolitik ▪ category-de:Minderheiten
rdfs:comment	▪ Minderheitensprachen bezeichnen die von ethnischen oder anderen Minderheiten verwendeten eigenständigen Sprachen.
rdfs:label	▪ Minderheitensprache
rdfs:seeAlso	▪ http://eu.dbpedia.org/resource/Hizkuntza_gutxitu ▪ http://pt.dbpedia.org/resource/Línguas_minoritárias
owl:sameAs	▪ http://ar.dbpedia.org/resource/لغة_أقليات ▪ http://ca.dbpedia.org/resource/Llengua_minoritària ▪ dbpedia:Minderheitensprache ▪ http://eo.dbpedia.org/resource/Minoritata_lingvo ▪ http://es.dbpedia.org/resource/Lengua_minoritaria ▪ http://fi.dbpedia.org/resource/Vähemmistökieli ▪ http://fr.dbpedia.org/resource/Langue_minoritaire ▪ http://gl.dbpedia.org/resource/Lingua_minoritaria ▪ http://hu.dbpedia.org/resource/Kisebbségi_nyelv ▪ http://it.dbpedia.org/resource/Lingua_minoritaria ▪ http://ja.dbpedia.org/resource/少数言語 ▪ http://lv.dbpedia.org/resource/Minoritātes_valoda ▪ http://nl.dbpedia.org/resource/Minderheidstaal
http://www.w3.org/ns/prov#wasDerivedFrom	▪ http://de.wikipedia.org/wiki/Minderheitensprache?oldid=102881455
foaf:isPrimaryTopicOf	▪ http://de.wikipedia.org/wiki/Minderheitensprache
is dbpedia-owl:officialLanguage of	▪ dbpedia-de:Schweden
is dbpedia-owl:wikiPageRedirects of	▪ dbpedia-de:Sprachliche_Minderheit
is foaf:primaryTopic of	▪ http://de.wikipedia.org/wiki/Minderheitensprache

Figure 8. Example output of the service for annotating web pages

4 Conclusions

This document presents the deliverable D3.1.2 Final text annotation prototype. First, the work in this deliverable consists in finding lexicalizations in all XLike languages of a set of resources from the DBpedia ontology in Wikipedia corpus. Multilingual information access can be facilitated by the availability of such lexicon, for example allowing for an easy mapping of natural language expressions in different languages to English ontology. Based on such cross-lingual lexicon, we build the final text annotation prototype using graph-based algorithm. The structure, functional specification and some details of technical specification of the final text annotation prototype are presented. Also, the definition of input, intermediary and output formats are given. In this deliverable no evaluation results have been included. The reasons for that are: (1) there is a deliverable dedicated to that topic (D7.3.2 Second benchmarking report) targeted for M24 when more processing material will be collected for testing; (2) the relevant golden standards for text annotation are still not available currently.

References

[D2.1.1] XLike deliverable “D2.1.1 – Shallow linguistic processing prototype”

[D3.1.1] XLike deliverable “D3.1.1 – Early Text Annotation Prototype”

[D3.2.1] XLike deliverable “D3.2.1 – Early ontological word-sense-disambiguation prototype”

[Milne2013] David Milne and Ian H. Witten. 2013. An open-source toolkit for mining Wikipedia. Artif. Intell. 194 (January 2013), 222-239.

[Bizer2009] Christian Bizer, Jens Leh3mann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: DBpedia - A crystallization point for the Web of Data. J. Web Sem. 7(3): 154-165 (2009)