Master's Thesis

Semantic Text Analysis and Interlinking

Web of Data: Novel approaches to media on the Internet

Ilir Ademi BSc

Semantic Text Analysis and Interlinking

Web of Data: Novel approaches to media on the Internet

Master's Thesis

at

Graz University of Technology

submitted by

Ilir Ademi

Institute for Information Systems and Computer Media (IICM),
Graz University of Technology
A-8010 Graz, Austria

25th Februar 2012

Advisor: Dipl-Ing. Dr. techn. Univ. -Doz. Denis Helic

Semantische Textanalyze und Interlinking

Web of Data: Neuartige Zugänge zu Medien im Internet

Diplomarbeit

an der

Technischen Universität Graz

vorgelegt von

Ilir Ademi

Institut für Informationssysteme und Computer Medien (IICM),
Technische Universität Graz
A-8010 Graz

25. Februar 2012

Diese Arbeit ist in englischer Sprache verfasst.

Begutachter: Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Abstract

Nowdays, an online user (editor) spends a lot of time invastigating upcoming topics, aggregating (online) content using information from various sources and merging all these small parts together shaping a fascinating story. Therefore, this master's thesis invastigates methods to produce editorial content quicker, enriched with multimedia objects and hyperlinks to advanced material using Linked Data. Unfortunately, a little has been said by the Linked Data community concerning the business potential of Linked Data and the challenges lying ahead for all enterprises, intending to use Linked Data in their professional production environments. Motivated from this lack of research and driven by the requirements of commercial content providers, the first version of SemTex (Semantic Text Analysis and Interlinking) project is developed and evaluated, aiming to support professional online editors in producing valuable online content, enriched with (multimedia) information from Linked Data sources.

Kurzfassung

Heutzutage verbrauchen die Editoren der Online-Medien einen beachtlichen Teil ihrer Arbeitszeit bei der Recherche in Internet bevor sie einen Beitrag schreiben. Dabei verwenden Sie verschiedenste verifizierte und unverifizierte Quellen, Links, Mediendaten wie Bilder oder Videos die entweder in internem Intranetz oder im Internet residieren. Dies macht mitttlerweile einen wichtigen Teil ihres Editorenalltags aus. Die meisten Vorgänge sind immer noch manuell auszuführen und hierbei handelt es sich nicht um die Kreativarbeit die eigentlich in Mittelpunkt stehen sollte. Diese Diplomarbeit setzt sich die Aufgabe die Vorgänge beim Suchen, Auffinden und Einbetten der begleitender Inhalte für Online- Beitragerstellung zu automatisieren und optimieren und somit den Rechercheaufwand zum Gunsten der Kreativität zu minimieren. Die Grundidee hierbei ist das Potenzial von Linked Data zu benutzen als vertraunswürdige Quelle um die gewünschte Effekte zu erzielen, sowie zu testen wie die Linked Data zu Lösung der Herausforderungen aus dem Umfeld der professionellen Produktionsumgebungen genutzt werden kann. Als Rahmen galten die Anforderungen der komerziellen Inhaltsvertreiber. Im rahmen der praktischen Arbeit wurde ein erster Prototyp von SemTex System basiert auf Linked Data Quellen und nativen NLP Methoden zu diesen Zwecken entwickelt und entsprechend evaluiert.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort Datum Unterschrift

Contents

List of Figures

List of Tables

Acknowledgements

My advisor Prof. Denis Helic deserves the foremost mention in these acknowledgements, because without him this master thesis would not have been possible. Like with many other students, Prof. Helic has been a patient, encouraging and supporting teacher, who helped me to bring out the best possible of my potential.

I especially wish to thank Wolfgang Halb from JOANNEUM RESEARCH, the project leader of this project, who was always available for questions and discussions whenever I knocked on his door.

Finally, a huge thanks to my parents who have taught me the value of a good education. Thanks to them, i had the opportunity to study at the technical university in Graz.

Last but not least, I am deeply indebted to my older brother for his support, patient and understanding during my studies.

Ilir Ademi
Graz, Austria, Februar 2012

Chapter 1

Introduction

> *" Unlike anything that has come before it, the combination of software and the World Wide Web has the potential to connect people and empower them in more ways than humanity has never seen. And it is possible to become immensely rich while moving humanity forward with the software that you create "*
>
> [Dare Obasanjo]

1.1 Motivation

Nowdays, the explosive growth of web technologies has drastically changed the online availability of data and the amount of electronically exchanged information. The collection of Web documents on the Internet doubles every six months. The Web has turned into a potentially reliable reference medium for many reasons, such as analysis, investigation and harvesting of information.

The people today spend most of the time using search engines such as Google, Bing or different online communities to discuss about new market trends, blogs(e.g. over Twitter), news(e.g. over BBC, Kleine Zeitung etc.) or to find friends (e.g. over Facebook). But finding the correct and important information from reliable sources of information over this massive information flood is one of most discussed topic. The most popular search services like Google and Yahoo use search based keyword machanism to find relavant keywords in the web documents. This means that only strings are compared in the text but the word meanings can not be detected.

Even if we now have our social networks (Web 2.0 platforms) which can provide selective access to information and personalized services which provide more or less exactly much information that we need is still to speak of "Information overload" of Internet.

With the transition from the Web 2.0 to the Semantic Web (Web 3.0) will help us finally to find it and get it what we are looking for: *The right information at the right time at right place.* The goal of web semantic, it is to upgrade the existing internet and computer application in order to make the information machine-readable and to link ideas, meanings and knowledge with each other. Semantic technologies enable to interpret information not only by human but also by machines.

In particular, the ability to link the semantic metadata with other open information could prove very valuable to e-commerce sector, companies in the world of media such as online newspapers, in the cultural sector and including operators of Web 2.0 platforms. Bloggers, CMS managers and online shop operators get a convenient way by providing additional information through automatic links from the Web of Data (WoD). For customers in the eCommerce it means that they must not strive only a search engine to get more information about a product. Consequently, they have a lot of information as a link in the product description. For Example; the Online sellers can enrich the information of their products by using additional information such as online field reports, customer statements, price comparisons with competitors or refer to additional products from other vendors. The users benefit from accurate and time-saving search for specific content and products, and the results can be due to extra information and better visualization to better understand and classify.

1.2 Problem Statement and Objectives

The scope of the currently available information on the web has to offer recently reached unimaginable scale, which directly correlates with an increasing number of users. However, offers of different platforms are hardly networked together. A network of such content, in particular taking into account semantic relationships can generate a great value for information providers; especially so because the respective content through various access points can be accessible. This is particularly important for companies in the media industry, as then if it comes to link together different sources and with varying quality demands - from professional information providers such as Network television and radio stations, media agencies, archives, museums or libraries, as well as by amateurs produced content (blogs, user generated content).

For example: an online editor usually investigates upcoming topics, aggregates (online) content from various sources, including user generated content taken from blogs, or professional content from press agencies, and merges all these small junks together shaping a fascinating story, capable of drawing the attention of the user.

Against this background, the project SemTex (Semantically Text Analasis and Interlinking) as novel approach to media on the Internet aims at creating a semantic and largely automatic networking of media information, taking into account of the user's specific context of production. Because of the economic importance for the media industry, this objective will be pursued for specific platforms with media information, however without leaving the feasibility of other applications.

The results of SemTex project are based on emerging technologies and latest trends. Thus, the implementation of the semantic network of SemTex is based on the currently highly topical Linking Open Data Initiative [1] which has the goal of making available the existing information sources (such as Wikipedia, Geonames, Eurostat etc) conform to Web of Data - the so-called datasets. In analogy will be applied the latest standards such as Friend of a Friend (FOAF) profiles for the description of user profiles, Dublin Core for the description of user items or the SIOC vocabulary (Semantically-Interlinked Online Communities). The advantages and the attractiveness of the so-modeled media content and user information is its automated ability to integrate into other internal and external resources of media providers (e.g. analogous to DBpedia). Furthermore, it can be integrated in provider's content management system.

Considering the scope of interests of a online content providers this thesis aims to achieve the following objectives:

- Automatic term extraction and classification, .
- Interlinking of entities (e.g. places, people, organisations) with existing datasets from Linked Data.
- Enriching of existing content of a provider with information from Linked Data.
- Visual analysis of multimedia information and extraction of entities.
- Publishing of enriched content in that manner that can interpreted by both humans and machines.

[1] LOD, http://linkeddata.org/

1.3 Thesis Outline

This thesis is structured in six chapters. The chapter 2 and chapter 3 belong to theorical parts. Chapter 2 focuses with methods of text analysis and terminology extraction. In the chapter 3 introduces the Semantic Web, technologies and its tools, including also the techniques of interlinking of data on the Web. Next, in the chapter 4 is concerned on related work on the semantic text structuring, analysis and interlinking. The subsequent chapter deals with SemTex application. Chapter 5 presents the concept, scenarios using Linked Data, the system architectural design, use cases, implementation ,evaluation and challenges of the SemTex application. Finally, the rest of the chapter 6 summarizes the conducted work and ideas for the future works.

Most parts of this work rely on the publication Halb et al. [2010] in Proceedings of 6th International Conference on Semantic Systems (I-SEMANTICS).

Chapter 2

Terminology extraction

“ Imitation is the sincerest flattery. ”

[Charles Caleb Colton, English writer, 1780–1832.]

This chapter gives the background of terminology extraction that is needed to understand a part of the work of this thesis. It includes the different NLP approaches and tools that can be apply to terminology extraction.

2.1 Introduction

The research on the definition and development of methods for extracting terms from texts assumed since the beginning a central role in the organization and harmonization of the knowledge enclosed in domain corpora, through the use of specific dictionaries and glossaries [Zanzotto, 2002]. Nowdays, the implementation of computational Natural Language Processing (NLP) approaches to term extraction, able to support and speed up the extraction process, lead to an increasing interest in using terminology also to build knowledge bases systems by considering information enclosed in textual documents. For this purpose, Ontology Learning and Semantic Web technologies often rely on domain knowledge automatically extracted from corpus through the use of tools able to recognize important concepts, and relations among them, in form of terms and terms relations.

In following sections we present a few, commonly agreed, linguistic and statistical approaches used in NLP to extract and recognize terms including the corresponding tools.

2.2 Term Extraction - Definition

Term extraction is an operation in language technology which takes as input a text document and produces as output a list of term candidates. Term candidates are words or phrases which are potential identified terms of the subject area represented by the input text document.

2.3 Terminology Extraction based NLP

Natural language processing (NLP) is a scientific field of computer science for analyzing and processing text of human (natural) languages that is based on both a set of theories and a set of technologies [Maynard et al., 2008]. NLP attempts to analyze and to process spoken or written language using natural language understanding and generation techniques. Many problems within NLP apply to both generation and understanding; for example, a computer must be able to model morphology (the structure of words) in order to understand the sentence, and a model of morphology is also needed for producing a grammatically correct sentence. In order to better understand these problems, it can be useful to identify the NLP approaches which fall roughly into two main categories: linguistic and statistical.

Linguistic approach

Linguistic approaches perform a deep analysis of linguistic phenomena using logic and/or rule-based models of natural language and are based on explicit representation of facts about language [Pazienza et al.]. In fact, the description of language analysis in the preceding task is given from a linguistic perspective. The main source of evidence in linguistic systems is a combination of human-developed rules and lexicons.

In logic based models, the linguistic structure is usually in the form of logic propositions where manipulations are defined by inference procedures. Rule-based models consist of a set of defined rules, an inference engine, and working space. Knowledge is represented according to the rules in the rule-base. The rule, whose condition is satisfied, is selected and executed from the inference engine.

Linguistic approaches can be used in variuos research fields and applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition.

Statistical approach

Statistical approaches apply stochastic, probabilistic and statistical methods and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge [Pazienza et al.]. These statistical models are mosly applied to complex natural language processing; for example longer sentences when processed with realistic grammars, yielding to thousands or millions of possible analyses. A frequently used statistical model is the accustic Hidden Markov Model (HMM). The technology for statistical approaches is part of machine learning and data mining, both of which are fields of artificial intelligence that involve learning from data.

In conclusion, term extraction can be solved using linguistic approaches, statistical approaches, or a mix of these two. An example of a term extraction method using linguistic method would first annotate text using a POS-tagger, then extract token sequences that apply common POS patterns for terms. A statistical method would e.g. calculate the inverse document frequency of a word and use this metric and a threshold to extract terms [Foo, 2009]. When extracting terms in a terminological setting, the result from a term extraction process are always called term candidates. Only after a manual validation process can they be called true terms.

2.4 SPPC

SPPC(Shallow Processing Production Center) is a high-performance domain-idependent extraction and navigation core system (see figure 2.1) of structured data from German free-text documents [Piskorski and Neumann, 2000]. SPPC consists of a set of domain independent shallow processing components among others tokenization, morphological analysis including online compound decomposition, part-of-speech filtering, named-entity recognition, sentence boundary detection, chunk and subclause recognition. It is capable of processing large collections of texts with high degree of robustness and efficiency (circa 30000 words per second in standard PC environment), since all components of the system were realized by means of cascaded optimized weighted finite state machines and generic dynamic tries. According to high linguistic coverage, the system has a good performance on all levels of processing.

The following components of SPPC were used for linguistic preprocessing:

- **Tokenizer** maps sequences of characters into greater units, called tokens and identifies the type

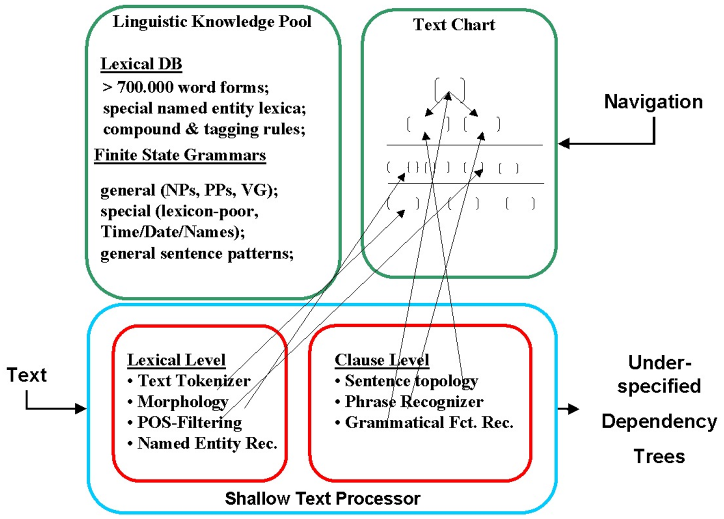


Figure 2.1: The system architecture of SPPC [Piskorski and Neumann, 2000].

of each token (e.g., two digit number, decimal number, lower case word, word begining with capital, candidate for abbreviation, number-word composition). There are over 50 default token types which simplifies the processing on higher stages (e.g., definition of named-entity recognition patterns).

- **Morphological Processor** processes each token which is identified as a potential word using lexical information and online recognition of compounds (e.g., Königreich - Kingdom). Morphological Processor uses a full-form lexicon containing currently over 700,000 entries.
- **POS(Part-of-Speech) Filtering** performs word-based disambiguation using three types of manually constructed filtering rules: case-sensitive rules, contextual filtering rules based on POS-information and rules for filtering out rare readings.
- **Named-entity Finder** finds entities (organizations, persons, locations), temporal expressions (time, date) and quantities (monetary values, percentages, numbers) using local pattern-matching techniques. For retrieving named-entities is used an additional lexicon for geographical names, first names (persons) and company names.
- **Chunk recognizer** extracts text fragments using divide and conquer parser.

2.5 KEA

Keyword extraction algorithm (KEA) is an algorithm for extracting keywords and keyphrases from text documents using machine-learning methods, especially Naive Bayes classifier [Uzun]. It can be either used for free indexing or for indexing with a controlled vocabulary. It has been developed by Witten et al. [1999], a research group for machine learning at the University of Waikato. Kea's extraction algorithm has two processes:

1. Training, which design a model for identifying keyphrases, using training documents where the author's keyphrases are known.
2. Extraction, selects keyphrases from a new document, using the above training model

In both processes will be choosen a set of candidate phrases from their input documents, and then calculate the values of certain attributes (or features) for each candidate. The Figure 2.2 depicts the Kea processes.

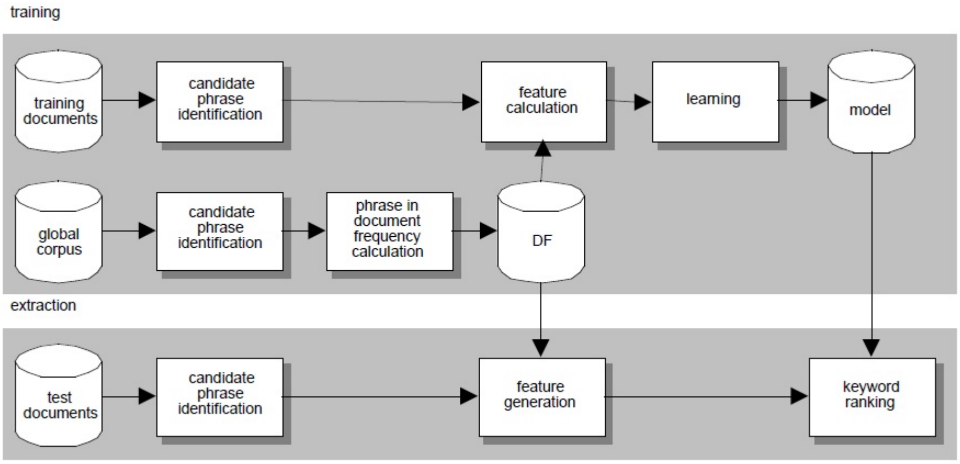

training
training documents
candidate phrase identification
feature calculation
leaming
model
global corpus
candidate phrase identification
phrase in document frequency calculation
DF
extraction
test documents
candidate phrase identification
feature generation
keyword ranking

Figure 2.2: The KEA training and extraction processes Witten et al. [1999]

Candidate phrases: Candidates are generated in three steps. It first cleans the input text such as punctuation marks, numbers, new lines and splits these sequences into tokens, then identifies candidates, and finally stems each candidate with the iterated Lovins stemmer [Lovins, 1968].

Features calculation: KEA calculates two features for each candidate: the TFxIDF measure (a measure of a phrase's frequency in the document collection) and first occurrence, which is the distance of the phrase's first appearance in the document.

Training: The training data will be created using Naive Bayes model, which consists of two sets of weights: for author's keyphrases and all other candidate phrases appearing in the document. Based on the training data the overall probability for each candidate being a keyphrase is calculated and ranked. The higher phrases ranked are included into the resulting keyphrase set.

2.6 Summary

This chapter gave us a short introduction of terminology extraction using two main approaches of NLP; linguistic and statistical. It also introduced SPPC and KEA, two mostly used tools that can be apply to extract german and english terms from free-text documents. The first part of this thesis has been focused on identification of related terms from the text which is implemented based on mentioned NLP methods and tools.

Chapter 3

Introducion to Semantic Technologies and Web of Data

> *" The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. "*
>
> [Tim Berners-Lee.]

3.1 Semantic Web Definition

In order to understand Semantic Web, we begin with by defining semantic. Semantic simply means meaning. Meaning enables a more effective use of the underlying data. Meaning is sometimes hard to find from most information sources, requiring users or complex programming instructions to supply it. For example, web pages contain information and associated tags. Most of these tags represent formatting instructions, such as *h1* to indicate a heading. Semantically, we know that words surrounded by *h1* tags are more important to the reader than other text because of the meaning of *h1*. Some of web pages add basic semantics for search engines using the *meta* tag; however, they are simply isolated keywords and lack linkages to provide a more meaningful context. These semantics are powerless and limit searches only to exact matches. Similarly, databases contain data and limited semantic hints, if well-named tables and columns surround the data.

Semantics give a keyword symbol useful meaning through the establishment of relationships. For example, a standalone keyword such as *building* exists on a web page devoted to ontologies. The *meta* tag surrounds the *building* keyword to indicate its importance. However, does *building* mean constructing an ontology or ontologies that focus on constructing buildings? The awkwardness of the previous sentence points out the difficulty in simply expressing semantics in English. Semantics are left for the human reader to interpret. However, if the keyword relates to other keywords in defined relationships, a web of data or context forms that reveals semantics. So *building* relates to various other keywords such as architect, building plans, construction site, and so on the relationships expose semantics.

If a formal standard captures the arrangement of terms, the terms adhere to specified grammar rules. It is even better if the terms themselves form an adopted standard or language. These two standards together, language and grammar, help to find the exact meaning, or semantics. As this contextual web of grammar rules and language terms expands through relationships, the semantics are further enriched.

Due to John Hebeler et al. [2009], the Semantic Web is defined as: "*The Semantic Web is simply a web of data described and linked in ways to establish context or semantics that adhere to defined grammar and language constructs*" .

3.2 The Semantic Web Architecture

As we have mentioned in the section above the architecture of semantic web will be strongly based on a hierarchy of languages. The Semantic Web was introduced by Tim Berners-Lee for the first time in one of his speeches in 1998 as an extension to the current web [Berners-Lee, September 1998], which include URI and Unicode. Tim Berners-Lee introduced the different versions of the Semantic Web architecture in 2000 [Berners-Lee, 2000] , 2003 [Berners-Lee, 2003], 2005 [Berners-Lee, 2005], 2006 [Berners-Lee, 2006]. Fensel is one of the main contributors in the Semantic Web area discussed the Semantic Web and the languages associated with its architecture in 2000 [Fensel, 2000], while in 2002, he described OIL and its relation to OWL and the future capabilities of OWL [Fensel, 2002]. Fensel was not the only scientist who made great efforts in this area, but there are Ian Horrocks [Horrocks et al., September 2005], Patel-Schneider [Horrocks et al., 2003] and Gerber [A.J.Gerber et al., 2007] also participated in this domain. The last modified version of semantic web architecture introduced by Tim Berners-Lee and refined by the mentioned contributers is depicted in figure 3.1. ed OIL and its relation to OWL and the future capabilities of OWL [Fensel, 2002]. Fensel was not the only scientist who made great efforts in this area, but there are Ian Horrocks [Horrocks et al., September 2005], Patel-Schneider [Horrocks et al., 2003] and Gerber [A.J.Gerber et al., 2007] also participated in this domain. The last modified version of semantic web architechture introduced by Tim Berners-Lee and refined by the mentioned contributers is depicted in figure 3.1.

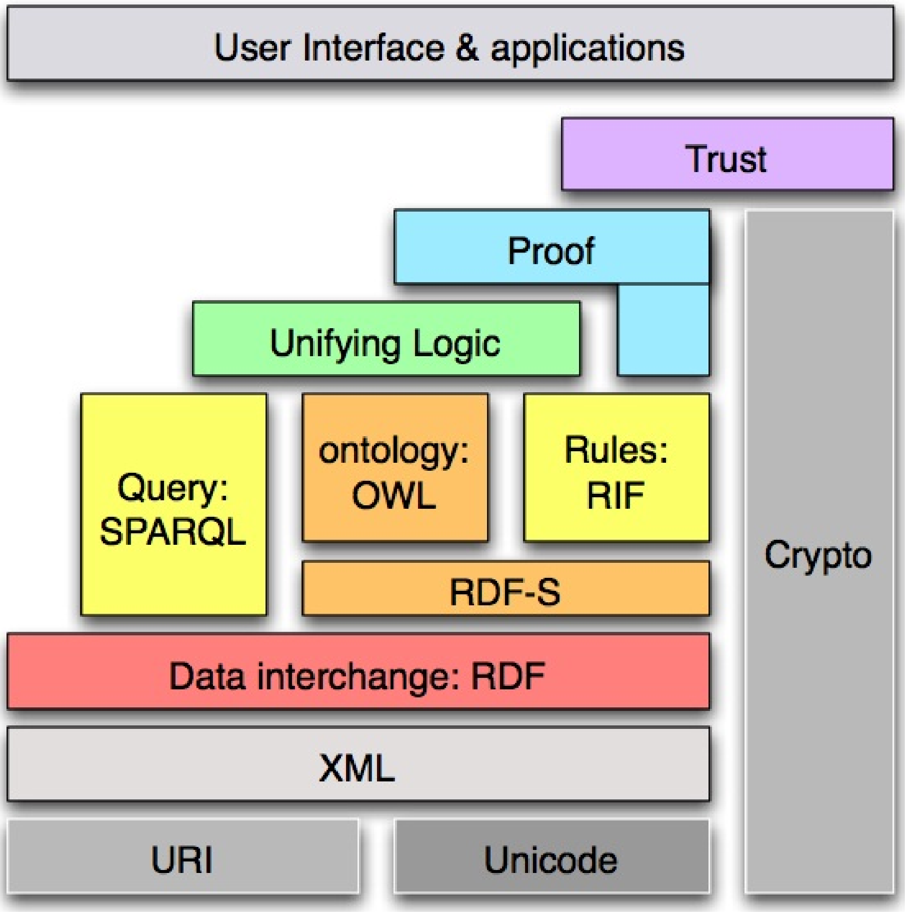


Figure 3.1: Latest version of the Semantic Web Stack [Horrocks et al., September 2005]

3.2.1 The Resource Description Framework (RDF)

RDF - the Resource Description Framework - is a model and language based on XML for representing information about resources in the World Wide Web of different types. It is particularly useful for storing metadata about shared resources.

Due to the W3C consortium: ”RDF is intended for situations in which this information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning. Since it is a common framework, application designers can leverage the availability of common RDF parsers and processing tools. The ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created ” [Manola and Miller, 2007]. With other words RDF emphasizes facilities to enable automated processing of Web resources.

However, by generalizing the concept of a ”Web resource”, RDF can also be used to represent information about things that can be identified on the Web, even when they cannot be directly retrieved on the Web. Examples; in resource discovery and classification to provide better search engine capabilities, in cataloging for describing the content and content relationships available at a particular Web site (such as social networks or online-communities by detecting the relationships between members) or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical ”document”; for describing intellectual property rights of Web pages; for expressing the privacy preferences of a user as well as the privacy policies of a web site and many others. In the future RDF with digital signatures will enhance Web Privacy with Policy Language and Trust which will be the key to building the ”Web of Trust” for e-commerce,collaboration, and other applications.

3.2.1.1 RDF Model

Information is represented as a set of assertions called statements made up of three parts: subject, predicate, and object. Because of these three parts, statements are also sometimes referred to as triples. The three elements of a statement have meanings that are analogous to their meanings in normal English grammar. The subject of a statement is the thing that statement describes, and the predicate describes a relationship between the subject and the object. In order to clarify this, consider the following Listing 3.1 about the workers of an organization.

```
Ilir knows Wolfgang.
Ilir’s surname is Ademi.
Wolfgang knows Selver.
Sandra works with Wolfgang.
```

Listing 3.1: Information about the workers

Figure 3.2 is a graphical representation of that small set of information. Assertions of this form naturally form a directed graph, with subjects and objects of each statement as nodes, and predicates as edges.

The nodes of an RDF graph are the subjects and the objects of the statements that make up the graph. There are two kinds of nodes: resources and literals. Literals represent concrete data values like numbers or strings and cannot be the subjects of statements, only the objects. A resource is simply a name that represents an object, act, or concept. Resource names take the form of Internationalized Resource Identifiers(IRI) [1].

[1] IRIs - http://tools.ietf.org/html/rfc3987

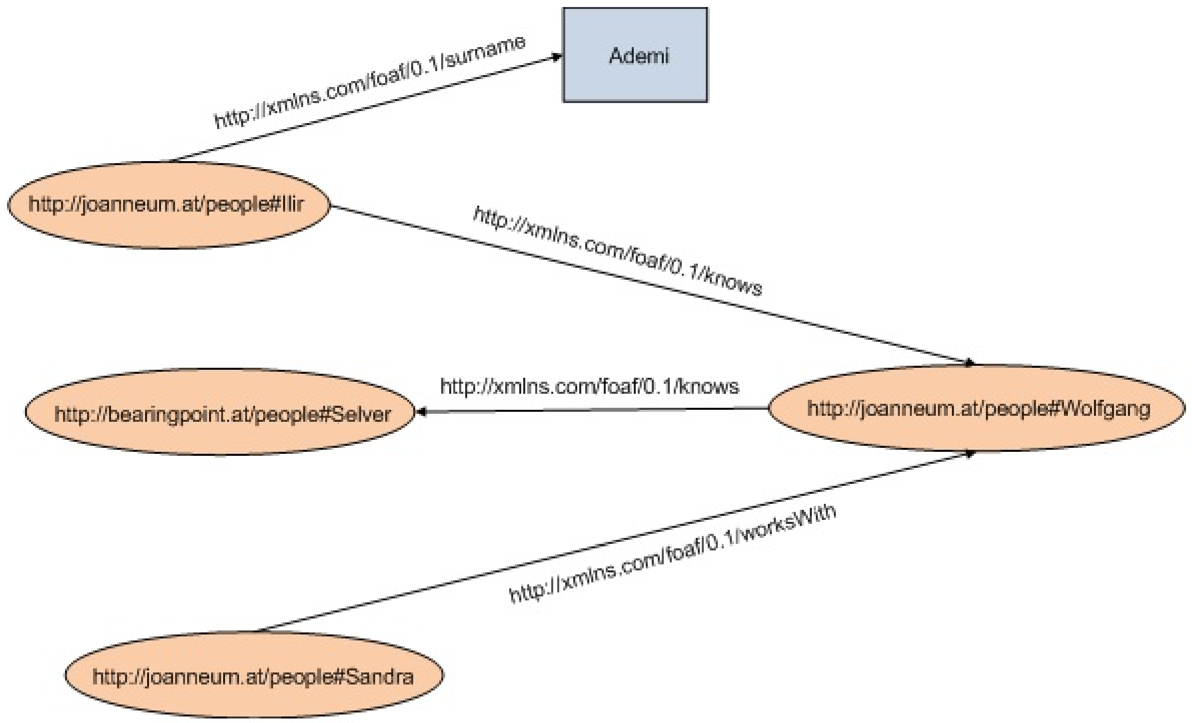

Ademi
http://xmlns.com/foaf/0.1/surname
http://joanneum.at/people#Ilir
http://xmlns.com/foaf/0.1/knows
http://bearingpoint.at/people#Selver
http://xmlns.com/foaf/0.1/knows
http://joanneum.at/people#Wolfgang
http://xmlns.com/foaf/0.1/worksWith
http://joanneum.at/people#Sandra

Figure 3.2: A graph representation of the sentences with IRIs from Listing 3.1

The Edges of an RDF graph are the predicates, also called properties, represent the connections between resources; predicates are themselves resources, however, and RDF statements can be made about predicates just as they can about any other resources. Like subjects, predicates are represented as IRIs.

3.2.1.2 RDF Serializations

In the previous section we have explored some of the characteristics of RDF, primarily the simple structure of the basic unit of RDF graphs, the graph structure of RDF, and the global namespace provided by the use of IRIs. RDF graphs are powerful tools for representing information, but they are abstract-good for human analysis but unsuitable for application exchange [John Hebeler et al., 2009]. Serialization makes RDF practical for information exchange by providing a way to convert between the abstract model and a concrete format, such as a file or other byte stream. There are several equally expressive serialization formats. Three of the most popular are RDF/XML, the Terse RDF Triple Language (Turtle), and N-Triples. All these serialization formats have special feature to represent these constructs more conveniently, but they all describe the same information.

RDF/XML : RDF/XML is an XML syntax defined by the W3C [2] for representing RDF triples, and it is the only standard exchange syntax for RDF serialization. This is the only syntax that is at least supported by RDF tools and all well-behaved Semantic Web applications. The only disadvantage of RDF/XML format is the comparison of two XML documents that look very different from each other may in fact be the same. The structure of RDF/XML graph can be shown in Listing 3.2.

As we can note from the Listing the whole content of RDF is contained within an rdf:RDF tag, which contains a series of rdf:Description elements. Another important item to note is the XML namespace declarations within the opening rdf:RDF tag. The RDF document that defines RDF itself can be found on the Web at http://www.w3.org/ 1999/02/22-rdf-syntax-ns.

[2] World Wide Web Consortium, http://www.w3.org/

```
<rdf:RDF
xmlns:people=”http://joanneum.at/people#”
xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
xmlns:foaf=”http://xmlns.com/foaf/0.1/”
xmlns:ext=”http://xmlns.com/foaf/0.1/
foaf-extension#”>
<!-- This is a comment. -->
<rdf:Description rdf:about=”http://joanneum.at/
people#Sandra”>
<ext:worksWith
rdf:resource=”http://joanneum.at/people#Wolfgang”/>
</rdf:Description>
<rdf:Description rdf:about=”http://joanneum.at/
people#Wolfgang”>
<foaf:knows
rdf:resource=”http://bearingpoint.at//people#Selver”/>
</rdf:Description>
<rdf:Description
rdf:about=”http://joanneum.at/people#Ilir”>
<foaf:surname>Ademi</foaf:surname>
<foaf:knows
rdf:resource=”http://joanneum.at/people#Wolfgang”/>
</rdf:Description>
</rdf:RDF>
```

Listing 3.2: The content from Listing 3.1, serialized as RDF/XML

Terse RDF Triple Language (Turtle): The Terse RDF Triple Language, or Turtle [3], is another serialization syntax for RDF.It allows RDF graphs to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. Therefore, Turtle is a more human-friendly and readable syntax. Turtle provides levels of compatibility with the existing N-Triples and Notation 3 formats as well as the triple pattern syntax of the SPARQL [4] W3C [5] Proposed Recommendation. Listing 3.3 shows the same graph used in Listing 3.1, this time serialized into Turtle.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix people: <http://joanneum.at/people/> .
@prefix ext: <http://xmlns.com/foaf/0.1/foafextension#>
.
# This is a comment.
people:Sandra ext:worksWith people:Wolfgang .
people:Wolfgang foaf:knows people:Selver .
people:Ilir
foaf:knows people:Wolfgang ;
foaf:surname ”Ademi” .
```

Listing 3.3: The content from Listing 3.1, serialized as RDF/XML

N-Triples: N-Triples is a simplified version of Turtle. It uses the same syntax for literal values, URIs, and comments but it does not support the @prefix directive or the ; or , shorthand for statements. A statement in N-Triples is a line-based, represented by a single line containing the sub-

[3] Turtle, http://www.w3.org/TeamSubmission/turtle/

[4] SPARQL, http://www.w3.org/TR/rdf-sparql-query/

[5] World Wide Web Consortium, http://www.w3.org/

ject, predicate, and object. N-Triples's simplicity can make it an attractive choice for serializing RDF, particularly in applications with streaming data. In Listing 3.4, the URI that in previous examples was http://joanneum.at/people# is replaced with a shorter URI, urn:sw:. The FOAF extension URI of http://xmlns.com/foaf/0.1/foaf-extension# is replaced with urn:swprg:foaf:, and the FOAF URI itself, http://xmlns.com/foaf/0.1/, is replaced with a shortened, print-friendly version, urn:foaf:.

```
<urn:sw:Sandra> <urn:swprg:foaf:worksWith> <urn:sw:Wolfgang> .
<urn:sw:Wolfgang> <urn:foaf:knows> <urn:sw:Selver> .
<urn:sw:Ilir> <urn:foaf:surname> ”Ademi” .
<urn:sw:Ilir> <urn:foaf:knows> <urn:sw:Wolfgang> .
```

Listing 3.4: The content from Listing 3.1, serialized as N-Triples

3.2.2 RDF Schema

The RDF data model does not allow to describe properties and relations between properties and nor does it define resources. Therefore, the RDF vocabulary description language [Brickley and Guha, 2004] has been introduced, various abbreviated as RDFS [6], RDF(S), RDF-S, or RDF/S. The first version was published by the W3C in April 1998, and the latest W3C recommendation was released in February 2004. RDF Schema provides a specific vocabulary for RDF that can be used to define taxonomies of classes and properties and simple domain and range specifications for properties. RDF Schema is a semantic extension which is written the same as RDF does. Using RDFS, you can arrange classes and properties in specialization and generalization hierarchies, define domain and range for properties, assert class membership, and specify and interpret datatypes. All resources in RDFS are considered members of the class of all RDF resources and as such are all instances. You can also describe those instances by making statements about them using properties or by explicitly making them members of other classes defined in an RDFS vocabulary. Listing 3.5 shows taxonomy of classes and properties and usage of range and domain of properties.

```
    @prefix :     <http://www.example.org/sample.rdfs#> .
    @prefix rdf:  <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .
    @prefix rdfs: <http://www.w3.org/2000/01/rdf−schema#>.

    :Cat       rdfs:subClassOf :Animal.
    :Person    rdfs:subClassOf :Animal.

    :hasChild rdfs:range :Animal;
              rdfs:domain :Animal.
    :hasSon   rdfs:subPropertyOf :hasChild.

    :Caddie   a :Cat.
    :Jason    a :Person.
    :Max      a :Person;
                :hasSon :Jason.
```

Listing 3.5: An Example of RDFS

[6] (RDF) Schema Specification 1.0 - http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

3.2.3 OWL

The Web Ontology Language (OWL)[McGuinness and van Harmelen, 2004] is an upgrade of the RDFS vocabulary with additional resources that can be used to build more expressive ontologies for the Web. OWL has added some restrictions regarding the structure and contents of RDF documents that make processing and reasoning more computationally decidable. OWL document contains the RDF and RDFS, XML Schema datatypes, and OWL namespaces. The OWL vocabulary itself is defined in the namespace http://www.w3.org/2002/07/owl# and is commonly referred to by the prefix owl. OWL2 [Group, 2009] is an refined version of original OWL vocabulary and reuses the same namespace. All namespaces used in an OWL document and their associated prefixes are listed in Table 3.1.

Table 3.1: Namespaces used in the OWL Web Ontology Language

Namespaces	Prefix
http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf
http://www.w3.org/2000/01/rdf-schema#	rdfs
http://www.w3.org/2001/XMLSchema#	xsd
http://www.w3.org/2002/07/owl#	owl

OWL is technically based on the RDF syntax and historically to DAML + OIL [7] (the name is the join of the names of the American proposal DAML-ONT [8], and the European language OIL [9])). The first OWL documents became a World Wide Web Consortium (W3C) recommendation on February 10th, 2004, after almost three years since the W3C Consortium started the so called ”Web Ontology Working Group”.

After a long time of development on October 27th 2009 became the second version of OWL (OWL2) a W3C Recommendation. OWL 2 is just an extension of original version of OWL with some new capabilities motivated by feedback from the users.

OWL 2 adds new functionality with respect to original OWL. Some of the new features are syntactic sugar while others offer new expressivity, including [Group, 2009]:

- keys,
- property chains,
- richer datatypes, data ranges
- qualified cardinality restrictions,
- asymmetric, reflexive, and disjoint properties and
- enhanced annotation capabilities.

3.2.3.1 Profiles of OWL

Both versions of OWL provide sublanguages or profiles of the language that give up some expressiveness in exchange for computational efficiency. The original OWL introduces three species each designed for specific area of appliance: OWL Full, OWL DL and OWL Lite.

[7] DAML + OIL - http://www.daml.org/2001/03/daml+oil-index.html

[8] http://www.daml.org/2000/10/daml-ont.html

[9] http://www.ontoknowledge.org/oil/

- ***OWL Full:*** OWL Full is not a sublanguage of OWL; rather, it is the full OWL language. It delivers the maximum expressiveness and the syntactic freedom of RDF. It is a pure extension of RDF with no computational guarantees. The classes in OWL Full are treated in the same time as a collection of individuals and as an individual in its own. Therefore, OWL Full is not decidable. There are still no algorithms that can produce all the entire entailment of semantics of a complex OWL Full knowledgebase [John Hebeler et al., 2009].

- ***OWL DL:*** OWL DL introduced some of restrictions on the use of OWL Full, including the separation of classes and individuals. These restrictions were designed to make OWL DL decidable and computable. The name of OWL DL is so given because it provides many of the capabilities of description logic (hence, OWL DL), a field of research that has studied the logics that form the formal foundation of OWL.

- ***OWL Lite:*** OWL Lite was essentially OWL DL with a subset of its language elements which was hoped to be a starting point for supporting the features of OWL Full. Unfortunately, OWL Lite is not widely used because it eliminates too many of the useful features of OWL Full without introducing enough of a computational benefit to make the reduced features attractive.

OWL 2 Profiles are sub-languages (syntactic subsets) of OWL 2 that offer some expressivity for better computational characteristics for tools and reasoners. Three different profiles are defined: OWL 2 EL, OWL 2 QL, and OWL 2 RL. These Profile are developed from the feedbacks from the users from online-communities and implementation technologies in mind. Each profile is based on OWL DL with some special restrictions.

- ***OWL 2 EL:*** The OWL QL profile provides algorithms for polynomial-time computation for all the standard reasoning tasks. The intention of this profile is to eliminate unnecessary features of OWL by providing only the expressive features of OWL that many existing large-scale ontologies (from various industries) require.

- ***OWL 2 QL:*** The OWL QL profile is designed to enable the satisfiability of conjunctive queries in logspace (*log(n)*) using standard relational database technology. It is ideal for users users who want to model the information contained in existing databases(e.g.; SQL) with relatively lightweight ontologies in order to better organize large numbers of individuals.

- ***OWL RL:*** The OWL RL profile is designed to be as expressive as possible while allowing implementation using rule-extended database technologies operating directly on RDF triples. It is ideal for users users who want to operate directly on data in the form of RDF triples with relatively lightweight ontologies in order to better organize large numbers of individuals.

3.2.4 The Rule Interchange Format (RIF)

The Rule Interchange Format (RIF [10]) activity within the World Wide Web Consortium (W3C) aims to develop a standard format for exchanging rules among diverse systems, especially on the Semantic Web. This format should works as an interlingua which allows the rules written for one application to be shared, published and re-used in other application or other rule engines. RIF includes three standard dialects, a Core dialect which is extended into a Basic Logic Dialect (BLD) and Production Rule Dialect (PRD)[Kifer, 2008]. The W3C initiated the RIF Working Group in 2005 with some ambitious and difficult goals [11]. Unfurtunaltely, none of the work produced by the RIF Working Group has yet been accepted as final recommendations by the W3C, and implementations have just begun to emerge.

[10] RIF - http://www.w3.org/2005/rules/wg/charter.html

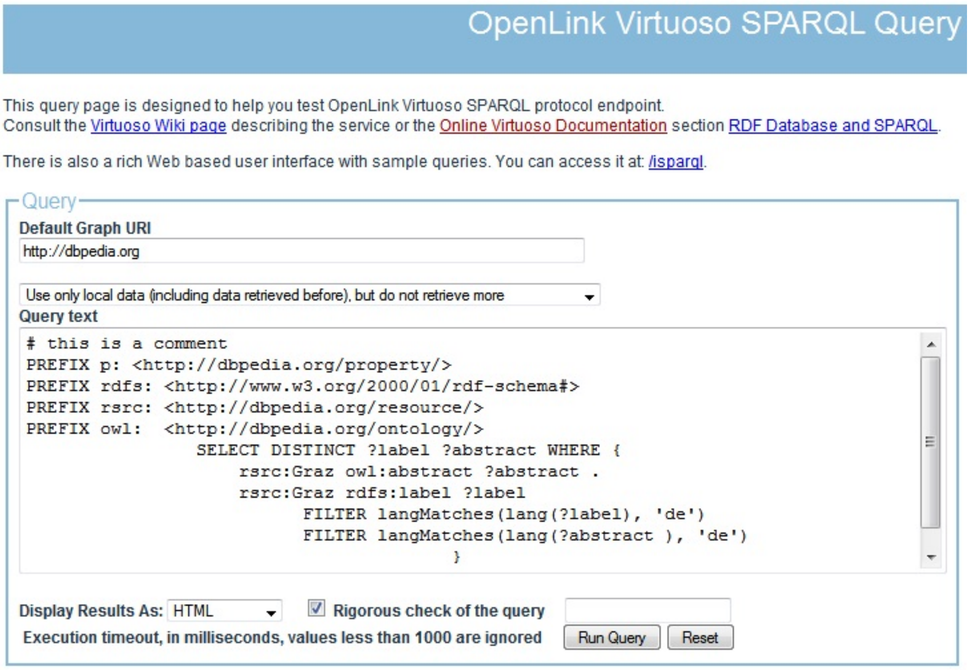
[11] RIF Goals - http://www.w3.org/TR/rif-ucr/#Goals

3.2.5 SPARQL

SPARQL [Prud'hommeaux and Seaborne, 2008] is a recursive acronym for SPARQL Protocol and RDF Query Language and is pronounced "sparkle". The SPARQL is considered by RDF Data Access Working Group (DAWG) of the W3C as key to semantic web as SQL is to a relational database , because of its W3C standardization, the wide community support, and the large number of publicly available endpoints. The SPARQL became an official W3C Recommendation on 15 January 2008. SPARQL allows applications to make sophisticated queries (query consisting of triple patterns, conjunctions, disjunctions, and optional patterns) against distributed RDF databases [Matthews, 2008].

The SPARQL endpoint [12] (or processor) is a service (although not necessarily a web service) that accepts and processes SPARQL queries and returns results in different formats depending on the query form. As we previously noted that SPARQL is both a query language and a protocol. Most of the people concentrate on the query language since it defines the syntax in which to create a query frame. The protocol is used to describe how a SPARQL client (such as one accessible via a web browser) which communicates to a SPARQL endpoint/processor (such as http://dbpedia.org/sparql) both in an abstract sense and using a concrete implementation based on WSDL 2.0 [John Hebeler et al., 2009].

In order to clarify the use of SPARQL the Figure 3.3 shows an example by using SPARQL web interface of DBPedia (http://dbpedia.org/sparql). The query asks for the abstract and label of the city Graz in german.



OpenLink Virtuoso SPARQL Query

This query page is designed to help you test OpenLink Virtuoso SPARQL protocol endpoint.
Consult the Virtuoso Wiki page describing the service or the Online Virtuoso Documentation section RDF Database and SPARQL.

There is also a rich Web based user interface with sample queries. You can access it at: /isparql.

Query

Default Graph URI: http://dbpedia.org

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

Query text

```
# this is a comment
PREFIX p: <http://dbpedia.org/property/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rsrc: <http://dbpedia.org/resource/>
PREFIX owl:  <http://dbpedia.org/ontology/>
                SELECT DISTINCT ?label ?abstract WHERE {
                    rsrc:Graz owl:abstract ?abstract .
                    rsrc:Graz rdfs:label ?label
                          FILTER langMatches(lang(?label), 'de')
                          FILTER langMatches(lang(?abstract ), 'de')
                                      }
```

Display Results As: HTML

☑ Rigorous check of the query

Execution timeout, in milliseconds, values less than 1000 are ignored

Figure 3.3: Querying DBpedia's SPARQL endpoint for the abstract and label of the city Graz in german

There are two main components to this query: the SELECT, WHERE clauses. The DISTINCT is a subcomponent of SELECT clause. The SELECT clause identifies which variables and their values will be returned from the query, the DISTINCT modifier eliminates duplicate solutions and the WHERE clause defines the graph pattern that will be matched against the data in DBpedia's RDF repository. *?label* and *?abstract* are variables representing either resources or data types. Variables are stated in lowercase letters, and keywords (such as PREFIX,SELECT,DISTINCT ,WHERE and FILTER) are capitalized. The langMatches and lang matches the language-range per the basic filtering scheme to find the german

[12] SPARQL protocol - http://www.w3.org/TR/rdf-sparql-protocol/

label and abstract for the keyword Graz. SPARQL supports the use of XML namespace prefixes with the PREFIX keyword (e.g.; PREFIX p: <http://dbpedia.org/property/>) by avoiding the long, repetitive namespaces in the query frame. The comments are written using the pound sign (#).

Table 3.2: Results for Graz

Label	Abstract
Graz	Graz ist die Landeshauptstadt der Steiermark ...

By default, DBpedia displays the results as an HTML table, as shown in Table 3.2. Returning to our example, we could view the results as XML,RDF/XML,JSON,N-Triples,Javascript and Spreadsheet instead of HTML.

3.3 Semantic Web Publishing

Publishing/sharing of information is the key of semantic web that gives us a powerful web platform. Information will be published on the web as documents with a set of XHTML code embedded that provides a structured set of data for computers to understand and even the meaning of the published information, making information search and data integration more efficient. The benefits of semantic publishing will enable the new semantic web - a web of data. This vision of the semantic web was described by Tim Berners-Lee in 1999 as:

"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web - the content, links, and transactions between people and computers. A "Semantic Web", which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize" [Berners-Lee, 2009].

Sharing of information is not only limited to SPARQL endpoint or by putting the static RDF document on your web server. But there are other ways of sharing information with others and building a chain of use and reuse from the data. These are new markup languages like RDFa and Microformats and other vocabularies.

3.3.1 RDFa

RDFa (or Resource Description Framework - in - attribute) is a W3C Recommendation that use the new XHTML attributes for embedding rich metadata in existing XHTML pages. The ultimate goal of RDFa is to make any RDF structure representable in pure XHTML which later can be extracted by compliant user agents. With a simple mapping it is possible to extract RDF triples from a RDFa annotated document.

RDFa is a syntax for expressing this structured data in XHTML. The rendered, hypertext data of XHTML is reused by the RDFa markup, so that publishers don't repeat themselves. The underlying abstract representation is RDF, which lets publishers build their own vocabulary, extend others, and evolve their vocabulary with maximal interoperability over time. The expressed structure is closely tied to the data, so that rendered data can be copied and pasted along with its relevant structure. [Adida and Birbeck, 2008].

RDFa supports the following attributes [Adida and Birbeck, 2008] listed below that can be used to carry metadata in an XML language:

- **about and src** - a resource URI or CURIE that is used to represent a subject in an RDF triple,
- **xmlns** - a prefix and qualified URL defining an XML namespace for the document,
- **rel and rev** - specify a relationship or reverse-relationship between two resources,
- **content** - optional attribute: specify an object in an RDF triple,
- **href and resource** - a resource URI or CURIE that is used to represent an object in an RDF triple with or without hyperlinks,
- **property** - specifying relationships between a subject and some literal text (also a "predicate"),
- **datatype** - optional attribute: a compart URI, or CURIE, that express a literal datatype,
- **typeof** - optional attribute: specify one or more RDF types that apply to the subject of the current triple,

RDFa has strong potential, especially with the backing of the W3C. It is unlikely that it will be the last technology for sharing semantics, as this burgeoning area has not seen the dust settle yet.

3.3.2 Microformats

”Designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards. Instead of throwing away what works today, microformats intend to solve simpler problems first by adapting to current behaviors and usage patterns (e.g. XHTML, blogging)”. [MF1, 2010]

Almost the same as RDFa the Microformats are simply XML tags that are embedded into XHTML web pages and support the declarative expression of semantics. The basic components of microformats are depicted in Figure 3.4 and are based on information located at www.microformats.org. The goal of microformats is to enrich the existing web pages with two additional attributes such as class and rel; it is easy for both humans and machines (such as intelligent agents) to specify the semantics that otherwise are not visible. Microformaty allows information intended for users (such as contact information, geographic coordinates, calendar events, and the like) to also be automatically extracted by softwares.

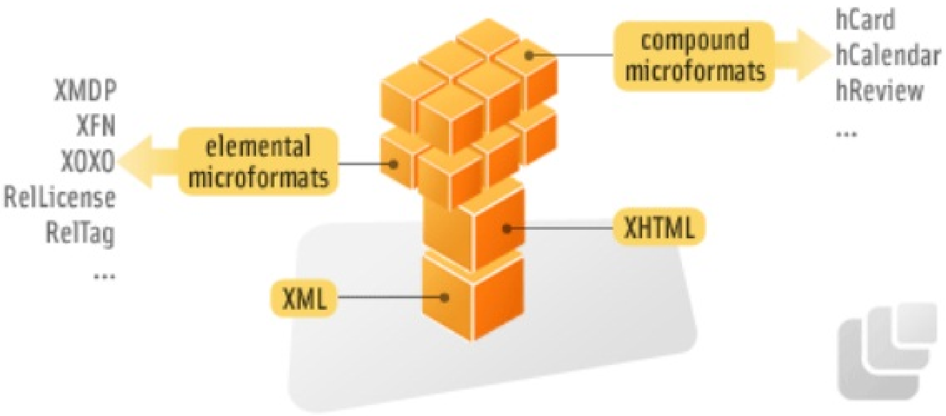


Figure 3.4: Basic components of microformats, source: www.microformat.org/about

For example: if Web page of John has a link of his friend Mathew, then we can add the attributes rel that expresses the relationship between John and Mathew such as rel = ”friend co-worker” (see Listing 3.6). The friend and co-worker values are defined as part of the XFN (XHTML Friends Network). microformat. As we can see from the example there is no need for recreating data scheme in a form such as OWL, or to conform to a one vocabulary because the rel tag is a standard XHMTL. Microformats were made to achieve some of of goals of semanting web using only the existing technologies of XHTML : structuring of web-based content, provides decentralized knowledge management and standards for community development to support structure reuse. Microformat sometimes are used to be called as the ”lowercase semantic web” since it is designed for human first and machine second.

```
  <div class=”vcard”>
    <div class=”fn”>John</div>
    <div class=”org”>An Example</div>
    <a class=”url” href = rel= ”friend co—worker”>Mathew’ page </a>
  </div>
```

Listing 3.6: An example of microformats

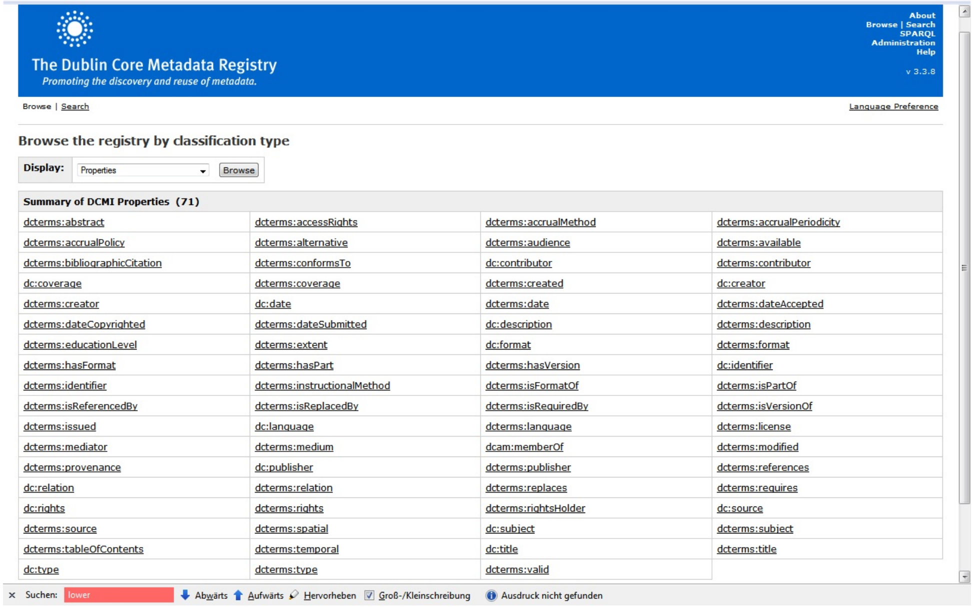
Microformats has proposed a set of standard specifications [MF1, 2010] that reuses XHTML attributes such as id and class to incorporate those specifications into XHTML documents:

- **hCard** [13] - a resource URI or CURIE that is used to represent a subject in an RDF triple,
- **hCalendar** [14] - a prefix and qualified URL defining an XML namespace for the document,
- **hReview** [15] - specify a relationship or reverse-relationship between two resources,
- **XFN** [16] - optional attribute: specify an object in an RDF triple,
- **XOXO** [17] - a resource URI or CURIE that is used to represent an object in an RDF
- **Rel-License** [18] - a resource URI or CURIE that is used to represent an object in an RDF

3.3.3 Ontologies and Vocabularies

3.3.3.1 Dublin Core

Dublin Core Metadata Initiative [19] is an open organization whose intent is to develop metadata standards for a broad range of applications, including document and multimedia description. The Dublin Core provides a registry of metadata terms (properties, classes, and other types of metadata terms) that can be displayed in Figure 3.5. The metadata terms are mainly pointed toward simple and common descriptive metadata, including title, type, description, authorship, and timestamp information, and are often used as annotation properties in ontologies.



Browse the registry by classification type

Display: Properties

Summary of DCMI Properties (71)			
dcterms:abstract	dcterms:accessRights	dcterms:accrualMethod	dcterms:accrualPeriodicity
dcterms:accrualPolicy	dcterms:alternative	dcterms:audience	dcterms:available
dcterms:bibliographicCitation	dcterms:conformsTo	dc:contributor	dcterms:contributor
dc:coverage	dcterms:coverage	dcterms:created	dc:creator
dcterms:creator	dc:date	dcterms:date	dcterms:dateAccepted
dcterms:dateCopyrighted	dcterms:dateSubmitted	dc:description	dcterms:description
dcterms:educationLevel	dcterms:extent	dc:format	dcterms:format
dcterms:hasFormat	dcterms:hasPart	dcterms:hasVersion	dc:identifier
dcterms:identifier	dcterms:instructionalMethod	dcterms:isFormatOf	dcterms:isPartOf
dcterms:isReferencedBy	dcterms:isReplacedBy	dcterms:isRequiredBy	dcterms:isVersionOf
dcterms:issued	dc:language	dcterms:language	dcterms:license
dcterms:mediator	dcterms:medium	dcam:memberOf	dcterms:modified
dcterms:provenance	dc:publisher	dcterms:publisher	dcterms:references
dc:relation	dcterms:relation	dcterms:replaces	dcterms:requires
dc:rights	dcterms:rights	dcterms:rightsHolder	dc:source
dcterms:source	dcterms:spatial	dc:subject	dcterms:subject
dcterms:tableOfContents	dcterms:temporal	dc:title	dcterms:title
dc:type	dcterms:type	dcterms:valid	

Figure 3.5: The Dublin Core Metadata Registry, source: http://dcmi.kc.tsukuba.ac.jp/dcregistry/

[13] hCard 1.0, http://microformats.org/wiki/hcard
[14] hCalendar 1.0, http://microformats.org/wiki/hcalendar
[15] hReview 0.3 , http://microformats.org/wiki/hreview
[16] XFN, http://microformats.org/wiki/xfn
[17] XOXO 1.0, http://microformats.org/wiki/xoxo
[18] Rel-Licence, http://microformats.org/wiki/rel-licence
[19] Dublin Core, http://dublincore.org

3.3.3.2 SKOS

Simple Knowledge Organization System (SKOS [20]) is an area of work developing specifications and standards that support the machanism for sharing data and linking knowledge organization systems via the Web. Many knowledge organization systems, such as thesauri, taxonomies, classification schemes and subject heading systems, share a similar concept, and are used in similar applications. SKOS provides a vocabulary to express matching of this similarity and makes it explicit, to enable data and technology sharing across various applications.

```
 <rdf:RDF
  xmlns:rdf=”http://www.w3.org/1999/02/22−rdf−syntax−ns#”
  xmlns:skos=”http://www.w3.org/2004/02/skos/core#”>

  <skos:Concept rdf:about=”http://www.example.com/concepts#animals”>
    <skos:prefLabel>animals</skos:prefLabel>
    <skos:altLabel>creatures</skos:altLabel>
    <skos:altLabel>fauna</skos:altLabel>
  </skos:Concept>
</rdf:RDF>
```

Listing 3.7: An example of SKOS, source: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/

The Listing 3.7 describe the concept animals with the labels animals,creatures and fauna which are considered synonyms of each other.

3.3.3.3 FOAF

The Friend of a Friend (FOAF [21]) is an ontology , which describes people, the relationship between them and the things they create on the World Wide Web. The ontology provides classes and properties for creating personal information, email addresses, online account and instant messaging information, as well as online documents and images. Listing 3.8 is a sample description of an individual using the FOAF ontology generated using the FOAF-a-Matic [22] web application.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf−schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .

    :me rdf:type foaf:Person ;
        foaf:family name ”Ademi” ;
        foaf:givenname ”Ilir” ;
        foaf:homepage <http://www.ilirnet.org> ;
        foaf:name ”Ilir Ademi” ;
        foaf:title ”Mr”.
```

Listing 3.8: An example of FOAF, generated with FOAF-a-Matic

[20] SKOS, http://www.w3.org/TR/skos-reference/

[21] FOAF, http://www.foaf-project.org

[22] FOAF-a-Matic, FOAF, http://www.ldodds.com/foaf/foaf-a-matic

3.3.3.4 SIOC

Semantically-Interlinked Online Communities Project (SIOC [23] - pronounced ”shock”) is a framework consisting of methods for describing and interconnecting of discussion methods such as web-blogs, wikis and mailing lists to each other by taking place on various online community forums. SIOC provides a Semantic Web ontology for representing rich data from the community sites in RDF. The SIOC Core Ontology are written using a computer language (RDF/OWL) because of its compatiblity with other existing ontologies such as Dublin Core and FOAF. Therefore, SIOC is often used in combination with the FOAF vocabulary for describing people and their friends, and the Simple Knowledge Organization System (SKOS) model for organising thesaurus like data, SIOC lets developers link user-generated content items to other related items, to people (via their associated user accounts), and to topics (using specifc ”tags”) [Bojars et al., 2008]. In 2004, the SIOC framework was started by John G. Breslin and Uldis Bojars at DERI, NUI Galway. The SIOC became in 2007 a W3C Member Submission.

```

@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://example.org/posts/post?id=1> a :Post;
  dcterms:created ”2009-10-06T08:34:29Z”;
  dcterms:title ”Sample SIOC post”;
  dcterms:subject ”Vienna”;

  :content ”Vienna is the capital of the Republic of Austria ...”;
  :has_container <http://example.org/posts/webblogs>;
  :has_creator <http://ilirademi.net/contact/>;
  :has_reply <http://example.org/posts/webblogs?id=22>;
  :topic [
    rdfs:label ”Vienna” ],
    [rdfs:label ”Capital City” ],
             <http://example.org/posts/category/Vienna/>,
       <http://example.org/posts/category/capital-city/>
    ].
```

Listing 3.9: An example of SIOC

The Listing 3.9 demonstrates an example of sioc document containig also the vocabulary of dublin core with the terms created, title and subject.

3.4 Semantic Web Frameworks

The Semantic Web is all about data on the web, useful semantic data. Reading or processing of these data require a machine, actually an application that interact with the formal structure of the Semantic Web. The Semantic Web is the brain or main storage of RDF data, whereas frameworks provide a useful programming body to apply the data of Semantic Web. Semantic Web frameworks are focused on object-oriented behaviors.

The processing of semantic data consists of some key areas [John Hebeler et al., 2009]:

- Managing and referencing accessible storage

[23] SIOC, http://sioc-project.org/

- Interrogating the Semantic Web data via navigation, search, and queries
- Populating or linking Semantic Web data to the referenced storage
- Reasoning via logic and rules across the Semantic Web data
- Adapting the framework to allow substitutions and customization for optimum results in a specific application domain

3.4.1 RDF Triple Stores

3.4.1.1 OpenLink Virtouso

OpenLink Virtuoso [24] is the first cross platform Universal Server (see Figure 3.6) that combines the the functionality of a traditional RDBMS, ORDBMS, virtual database, RDF, XML, free-text, web application server and file server functionality in a single server solution. The OpenLink Virtuoso has been developed by Kingsley Idehen and Orri Erling as the chief software architects. The goal of OpenLink Virtuoso was to design a unique system of threading support and multiple CPUs that reduces the cost of bringing together data from different data sources with the view to accelerating the production of information by your Query Tools, Web & Internet Application Development Environments, Traditional Application Development Tools, and Desktop Productivity Tools. In Virtuoso, RDF data can be stored as RDF quads [Erling, 2008] , i.e. graph, subject, predicate, object tuples. All such quads are in one table, which may have different indexing depending on the expected query load. Virtuoso supports embedding SPARQL into SQL as a method of accessing RDF triples that is stored in the knowledgebase. It also provides the SPARQL endpoint to perform SPARQL queries and uploading of data over HTTP from various large data sources such as DBpedia.

3.4.1.2 Jena

Jena[25] is an open source framework for building semantic web applications in java programming language which is grown out of work with the HP Labs Semantic Web Programme. It provides a programmatic environment that allows different storage implementations to be used with the common Jena APIs for RDF, ontologies and SPARQL query and includes a rule-based inference engine. The class *Model* of Jena framework is an API for dealing with a set of RDF triples that can be created from a remote file or filesystem.

SDB Jena[26]: SDB is a Jena's framework that provides for scalable storage and query of RDF datasets using conventional SQL databases for use in standalone applications, J2EE and other application frameworks.SDB is designed specifically to support SPARQL.

TDB Jena[27]: TDB is a component of Jena. It provides for large scale storage and query of RDF datasets using a Java engine. TDB supports SPARQL and provide provide the storage layer for both a single machine usage and also distributed clusters of industry. standard servers.

Joseki[28]: Joseki is a web server that support SPARQL protocol and the SPARQL RDF Query language. It enable SDB store and TDB store to be queried over HTTP.

ARQ [29]: ARQ is a query engine for Jena that supports the SPARQL RDF Query language. The storage engines of ARQ are SDB and TDB Jena. ARQ supports the following features:

[24] OpenLink Virtuoso, http://virtuoso.openlinksw.com/

[25] Jena, http://openjena.org/

[26] SDB Jena, http://openjena.org/SDB/

[27] TDB Jena, http://openjena.org/TDB/

[28] Joseki, http://www.joseki.org/

[29] ARQ, http://jena.sourceforge.net/ARQ/

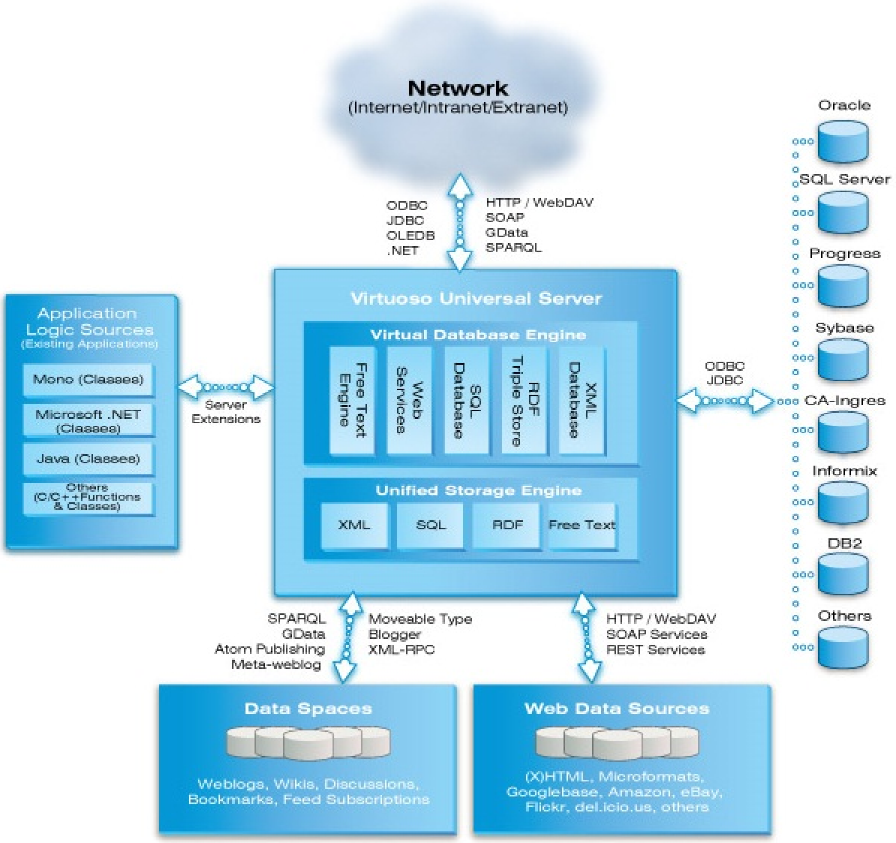


Figure 3.6: The architecture of OpenLink Virtuoso

- Standard SPARQL
- SPARQL/Update
- Free text search via Lucene
- Access and extension of the SPARQL algebra
- Client-support for remote access to any SPARQL endpoint
- Support for custom filter functions
- Support for federated query
- Aggregation, GROUP BY and assignment as SPARQL extensions
- Property functions for custom processing of semantic relationships

3.4.1.3 Sesame

Sesame [30] is widely used RDF framework and server. It is written in java which is fully extensible and configurable with respect to storage mechanisms, inferencers, RDF file formats, query result formats and query languages. Sesame can be easily deployed as a servlet in a servlet container such as Apache Tomcat. Thus, Sesame offers a JBDC-like user API, streamlined system APIs and a RESTful HTTP interface supporting the SPARQL Protocol for RDF. Sesame was firstly developed by Aduna (then known

[30] Sesame, http://www.openrdf.org/

as Administrator) as a research prototype for the EU research project On-To-Knowledge until 2001. Then the development of Aduna has continued with with NLnet Foundation, developers from Ontotext, and some other users who help to fix the bugs and other improvements. Main developers of Sesame are Arjohn Kampman and Jeen Broekstra.

3.4.1.4 D2D Server

D2R Server is mostly used as a tool for publishing the content of relational databases on the Web using SPARQL-Endpoints, a global information space consisting of linked data. The SPARQL-Endpoint enables applications to search and query the database using the SPARQL query language over the HTTP protocol. Information on the Web is structured and represented in a form of RDF. DR2 Server allows RDF data to be searched and browsed by using a customizable D2RQ mapping [Bizer and Cyganiak, 2009] that maps the content of relational databases into the format of RDF . The data browsing and searching are the main access paradigms to the Web of Data (see Figure 3.7). Requests from the Web are translated on-the-fly into SQL queries via D2RQ mapping. This allows automatically to publish the RDF data from the large relational databases and simultaneously avoid the need for replicating the data into RDF triple store.

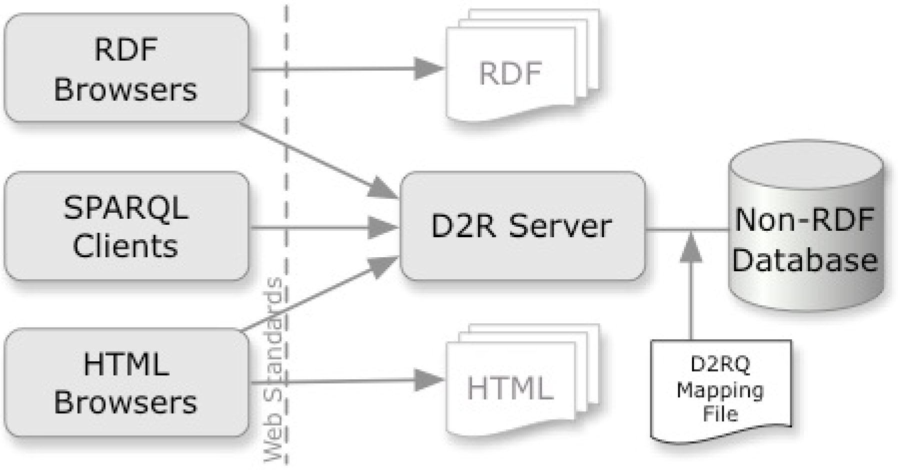


Figure 3.7: A Model of D2D Server

3.4.2 Semantic Reasoners

As discussed before, we use OWL to enrich information by definging the semantics. In order to interpret the semantics and enrich information, an inteference machanism is applied. Applications that apply inference are piece of software, often refered to as semantic reasoner, reasoning engine, rules engine, or simply a reasoner. A reasoner is able to infer new information based on the contents of a knowledgebase by using rules and a rule engine, triggers on RDF store, decision trees, tableau algorithms, or even programmatically using hard coded business logic. Many reasoners use first-order predicate logic of executing inference in rule-based reasoning; forward chaining and backward chaining. The combination of these two methods is referred to as hybrid reasoners which is applied by many semantic engines.

In addition, reasoners introduce two different types: internal reasoners built into the Jena framework, external reasoners offered as external Java files or remotely offered via the DL Information Group [31] (DIG) interface.

[31] DIG, http://dig.cs.manchester.ac.uk/overview.html

Jena Rules or Jena2 inference subsystem[32] is an internal reasoner with a range of inference engines or reasoners built in Jena framework. Such engines derive supplementary RDF assertions from some base RDF together with any optional ontology information and the axioms and rules associated with the reasoner. The mechanism of Jena Rules is primarly aimed to support the use of languages such as RDFS and OWL which allow supplementary facts to be inferred from instance data and class descriptions. Jena reasoner may create model to associate a data set with some reasoner. The rules of the model fires in accordance to its configuration. A reasoner may fire in a forward-chaining mode, backward-chaining mode, or a hybrid of both. Forward rules fire whenever a new rule is added to the rule reasoner or new statements are added to the associated model. Backward rules fire whenever a query is executed on the associated model. A rule may add new statements to the model and new rules to the rule reasoner. Thus a rule may create another rule and so on. A structure of Jena rule provides a list of body terms or premises (the if clause) and a list of head terms or conclusions (the then clause). The listing 3.10 is an example of Jena Rules; a male driver living in New York or Vancouver and having less than 2 accidents is eligible for insurance.

```
<ex:Driver rdf:about=”http://example.com/John”>
  <ex:state>Vancouver</ex:state>
  <ex:accidentsNumber rdf:datatype=”http://www.w3.org/2001/XMLSchema\#integer”>1</
      ex:accidentsNumber>
</ex:Driver>

@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns\#
@prefix ex: http://example.com/
[eligibleDriver_1: (?d rdf:type ex:EligibleDriver)
                   <-
                   (?d rdf:type ex:Driver)
                   (?d ex:state ”New York”)
                   (?d ex:accidentsNumber ?an)
                   lessThan(?an,2)]

[eligibleDriver_2: (?d rdf:type ex:EligibleDriver)
                   <-
                   (?d rdf:type ex:Driver)
                   (?d ex:state ”Vancouver”)
                   (?d ex:accidentsNumber ?an)
                   lessThan(?an,2)]
```

Listing 3.10: An example of Jena Rules. Source: http://hydrogen.informatik.tu-cottbus.de

Pellet

Pellet[33] is an Java-based OWL DL reasoning engine that supports a majority of the constructs of OWL, including those introduced in OWL 2. It is based on the tableaux algorithms developed for expressive Description Logics. It also can be easily integrated in both Jena and RDF/OWL toolkits and also provides a DIG interface. Pellet is developed and commercially supported at Clark and Parsia LLC. Mike Smith is one of the primary developers of Pellet. In addition, Pellet implements most of the state of the art optimization techniques provided in the DL literature including Normalization, Simplification, Absorption, Semantic Branching, Backjumping,Caching Satisfiability Status, Top-Bottom Search for Classification, and Model Merging [Sirin et al., 2007].

Pellet provide a number of features and capabilities either driven by OWL requirements or Semantic Web issues:

[32] Jena2 , http://jena.sourceforge.net/inference/

[33] Pellet, http://www.mindswap.org/2003/pellet/

- Ontology analysis and repair,
- Entailment,
- Conjunctive ABox query,
- Datatype Reasoning,
- User-defined Simple Datatypes,
- Multi-Ontology Reasoning using E-Connections and
- Ontology Debugging

Currently, Pellet uses distinct existing applications and there is some ongoing work to integrate Pellet to different:

- Ontology Development (such as: Protege and Swoop),
- Web Service Composition and
- Rule integration

3.5 Interlinking of Data on the Web

Until now we have have discussed all the methods of creating repositories and registries[John Hebeler et al., 2009] for the existing data on the Web. Another approach is the publishing of data and distributing the wealth of ontologies that allows the user to automatically interlink them together. This was a major step forward in realising Berners-Lee, Handler and Lassila's original vision of the web that can "understand and satisfy the requests of people and machines to use the web content" - i.e. the Semantic Web [Berners-Lee et al., 2001].

Data on the Web from distinct data sources uses typed links to be linked with each other which is refered to "Linked Data" [Bizer et al., 2009]. This can be an posted article to different Web Platforms which contains entities such as Graz. Therefore, Graz can be identify using the RDF links from DBPedia. (http://dbpedia.org/resource/Graz),Geonames (http://sws.geonames.org/2778067/) and other open datasets. Practically, Linked Data refers to data published on the Web, which are machine readable, can be linked to other external data sets and can in turn be linked to from external data sets. Berners-Lee (2006) outlined four key rules for publishing data on the Web as a single global data space:

- All things should be identified using URIs
- Use HTTP URIs, so that people can look up over HTTP those names for detailed information.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other related URIs using owl:sameas, so that they can discover more things.

These rules are the basic principles of Linked Data and provide a recipe for publishing and interlinking data using the infrastructure of the Web while relying to its architecture and standards.

3.5.1 Linked Open Data on the Web

Based on the Linked Data [34] principles has arised the idea of creating connected datasets each other which contain information in RDF format. In January 2007 has started the project of Linked Open Data (LOD) with the support of W3C Semantic Web Education and Outreach Group with the objective of bootstraping of the Web of Data by wrappers around relational databases or using APIs that are available under open licenses, and RDFizing them according to the Linked Data principles, interlinking and publishing them on the Web.

In early 2007 were only a modest number of datasets, used for farther research and development in universities and small companies (e.g. DBPedia [35] and Geonames [36]). Since that time the number of datasets is still growing by involving of large databases such as BBC, Thomson Reuters, digital libraries and etc. Everyone can publish his datasets on the LOD project considering the Linked Data principles and interlinking it with other existing datasets.

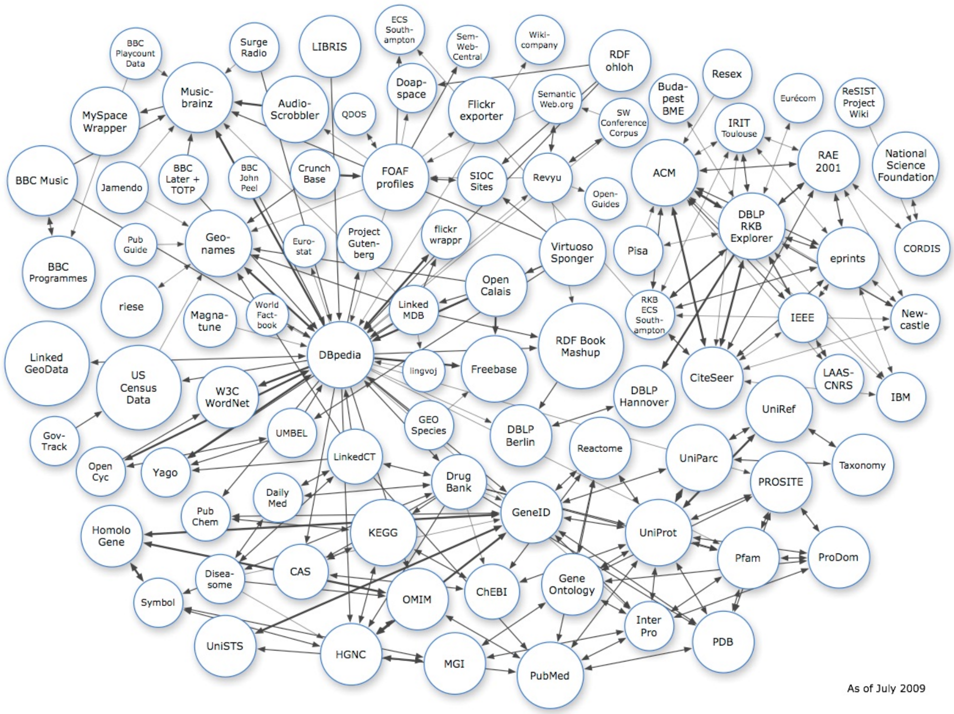


Figure 3.8: Linking Open Data (LOD) Project Cloud Diagram, Source: http://linkeddata.org/home

The figure 3.8 shows the LOD cloud diagram representing the web of distinct datasets of different nature, as of June 2009. Arcs indicate links between the items of two datasets, that can bidirectional. Heavier arcs represents roughly the number of links between datasets.

The datasets in LOD cloud are from diverse nature, containing data about people,companies, geografic locations, books [Bizer et al., 2007b], movies [Hassanzadeh and Consens, 2009], scientific publications (Van de Sompel et al., 2009), music, television and radio programmes [Kobilarov et al., 2009], online communities, statistical data, census results, and reviews [Heath and Motta, 2008], genes, proteins, drugs and clinical trials [Bellea et al., 2008],[Jentzsch et al., 2009].

[34] Linked Data, http://linkeddata.org/

[35] DBPedia, http://dbpedia.org/

[36] Geonames, http://www.geonames.org/

3.5.2 SILK - An Interlinking Framework for the Web of Data

Silk[37] is an interlinking framework which identifies the relationship and set RDF links between data items within distinct datasets of LOD cloud in order to create the Web of Data. Silk uses a declarative language, called Silk Link Specification Language (Silk-LSL). Silk-LSL specifies which type of RDF links should be identify between two distinct datasets and the interlinking conditions that must be fulfilled in order generate the set of RDF links. The interlinking conditions use different similarity metrics such as: jaro, jaro Winkler, taxonomy, qGram, String and URI compare. These similarity results can be weighted and combined all together using various similarity aggregation functions such as; AVG - weighted average, MAX - choose the highest value, MIN - choose the lowest value, EUCLID - Euclidian distance metric and PRODUCT - weighted product. Data sources can be accessed using either a local or a remote SPARQL Endpoints. Silk is implemented in Python language and is run as a batch process on the command line.

The main features of the Silk framework are as follows [Volza et al., 2009]:

- it supports the generation of owl:sameAs links as well as other types of RDF links,
- it provides a flexible, declarative language for specifying link conditions,
- it can be employed in distributed environments without having to to replicate datasets locally,
- it can be used in situations where terms from different vocabularies are mixed and where no consistent RDFS or OWL schemata exist.
- Silk implements various caching, indexing and comparison preselection methods to increase performance and reduce network load.

The listing 3.11 illustrates an example of Silk-LSL construct. In this case, we want to identify owl:SameAs links between the URIs that are used by DBpedia and by GeoNames to identify cities according to Silk restrictions and specifications.

3.5.3 OOD-Linker

ODD-Linker is an interlinking framework implemented on top of a mining techniques for term matching in relational databases. ODD-Linker uses SQL queries for discovering and comparing resources. The framework uses as input format a dedicated LinQL in order to translates link specifications into such SQL queries. ODD-Linker is particularly designed to be used with relational databases exported in RDF. The framework was used to interlink the Linked Internet Movie Database to DBPedia [Hassanzadeh and Consens, 2009].

3.5.4 Knofuss

The Knofuss is another framework of data interlinking providing support for data-level integration of ontological data (also called knowledge fusion)[Scharffe and Euzenat, 2009]. It uses existing ontology alignment specifications for resources comparison, as well as the comparison methods such as string matching methods or similarity metrics. When the two datasets to interlink are structured using different ontologies, an ontology alignment can be specified in the ontology alignment format [euzenat:2004], allowing to reuse one of the specified ontology matching systems available. The results are generated using SPARQL queries that are targeted to the datasets to interlink. Knofuss being originally designed for data fusion, it is possible to specify a fusion strategy. A post-processing step is performed by the tool in order to verify the consistency of the dataset resulting from the fusion operation[**?**]. This tool works with local copies of the datasets and is implemented in Java.

[37] Silk, http://www4.wiwiss.fu-berlin.de/bizer/silk/

```
<Silk>

  <Prefix id=”rdfs” namespace=”http://www.w3.org/2000/01/rdf−schema#” />
  <Prefix id=”dbpedia” namespace=”http://dbpedia.org/ontology/” />
  <Prefix id=”gn” namespace=”http://www.geonames.org/ontology#” />

  <DataSource id=”dbpedia”>
    <EndpointURI>http://demo_sparql_server1/sparql</EndpointURI>
    <Graph>http://dbpedia.org</Graph>
  </DataSource>

  <DataSource id=”geonames”>
    <EndpointURI>http://demo_sparql_server2/sparql</EndpointURI>
    <Graph>http://sws.geonames.org/</Graph>
  </DataSource>

  <Interlink id=”cities”>
    <LinkType>owl:sameAs</LinkType>

    <SourceDataset dataSource=”dbpedia” var=”a”>
      <RestrictTo>
         ?a rdf:type dbpedia:City
      </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource=”geonames” var=”b”>
      <RestrictTo>
        ?b rdf:type gn:P
      </RestrictTo>
    </TargetDataset>

    <LinkCondition>
      <AVG>
        <Compare metric=”jaroSimilarity”>
          <Param name=”str1” path=”?a/rdfs:label” />
          <Param name=”str2” path=”?b/gn:name” />
        </Compare>
        <Compare metric=”numSimilarity”>
          <Param name=”num1” path=”?city1/dbpedia:populationTotal” />
          <Param name=”num2” path=”?city2/gn:population” />
        </Compare>
      </AVG>
    </LinkCondition>

    <Thresholds accept=”0.9” verify=”0.7” />
    <Output acceptedLinks=”accepted_links.n3” verifyLinks=”verify_links.n3” mode=”
        truncate” />
  </Interlink>

</Silk>
```

Listing 3.11: An example of Silk construct. Source: http://www4.wiwiss.fu-berlin.de/bizer/silk/

3.6 Summary

This chapter presents the fundamentals of the Semantic Web including architecture, technologies, languages and frameworks for creating repositories and registries for the data on the Web. Then, it indroduces the methods and tools for publishing of data and distributing the wealth of ontologies in Linked Open Data cloud. Linked Open Data allows the users or machines to automatically consume and (inter)link its information which is also the aim of this work. The new interlinking methods are indroduced to interlink concepts between datasets in Linked Open Data cloud such as SILK OOD-Linker and Knofuss.

Chapter 4

Existing Works on Semantic Text Structuring, Analysis, and Interlinking

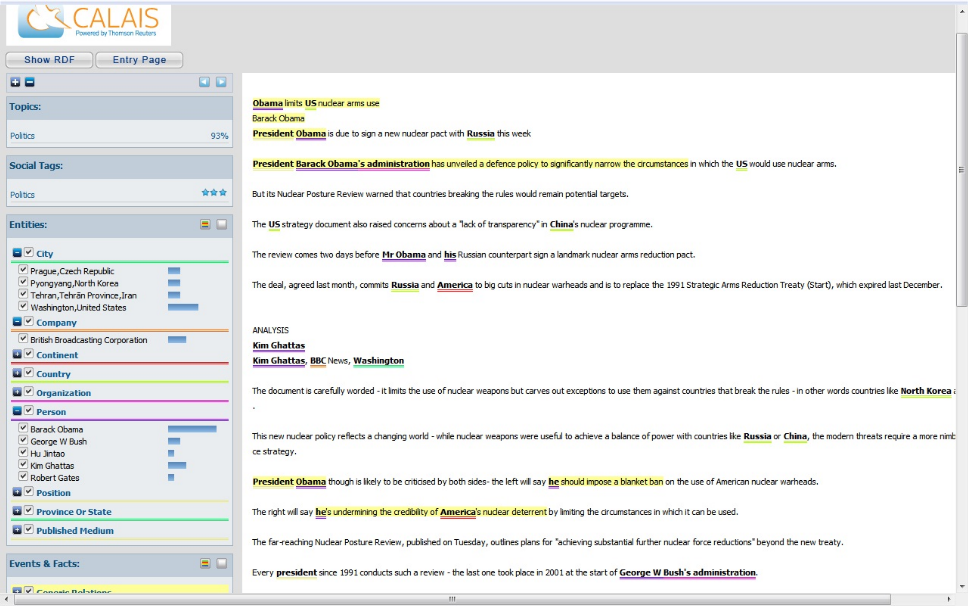
> *“ ... fighting the web is like holding back the ocean; it will route around you or it will wear you down, but will never go away, and it will never tire or give up. ”*
>
> [DeWitt Clinton, On Fighting the Web Itself.]

4.1 OpenCalais

The OpenSource Initiative ”OpenCalais” of Thomson Reuters is about facilitating the semantic knowledge representation of the content on the World Wide Web and bring it forward. At this point will be presented the Annotation Service ”OpenCalais Web Services” as an available semi-automated Annotationsinstrument. The OpenCalais Web Service is an API that accepts unstructured text, processes it using Natural Language Processing (NLP) and Machine Learning (ML) algorithms, returns RDF-formatted entities, facts and events and is used to fuel the applications that build the Calais initiative. The semantic enrichment of the data takes place by means of RDF. For the transmission and exchange of metadata, the Calais uses different search engines or websites.

OpenCalais Web Services are aimed mainly at programmers and website operators and represent different plug-ins available for free. Calais Viewer is the best way to get a quick glimpse of OpenCalais. We can manually try the web service by typing in or pasting text into the viewer box, without needing an API key. For example, we used a unstructured news release and got the results depicted in Figure 4.1. It identifies the correct topic of the article - Politics, the City that the article is about - Beijin in China, Washington in United States and etc., three Person entities: Barack Obama, Greg Mello, Joseph Cirincione and etc. In the Toolbar, the annotation in RDF format can be displayed.

The semantically annotated content can now be only linked to freely-sources or are on their own databases and stand available on their applications. Furthermore, the OpenCalais provides an extension for Firefox and Internet Explorer called Gnosis (see 4.2). Gnosis is a browser extension that automatically analyzes content as you browse and provides a variety of tools to explore the people, places, companies, and other items that you’re reading about. Available information on these categories will be displayed at the ”roll over” with the mouse. In context menu, the interested users will be redirected to other sources such as: Google, Wikipedia, Technorati or Google-Maps.



Topics:
Politics 93%

Social Tags:
Politics ★★★

Entities:
☑ City
☑ Prague,Czech Republic
☑ Pyongyang,North Korea
☑ Tehran,Tehrān Province,Iran
☑ Washington,United States
☑ Company
☑ British Broadcasting Corporation
☑ Continent
☑ Country
☑ Organization
☑ Person
☑ Barack Obama
☑ George W Bush
☑ Hu Jintao
☑ Kim Ghattas
☑ Robert Gates
☑ Position
☑ Province Or State
☑ Published Medium

Events & Facts:
☑ Generic Relations

Obama limits **US** nuclear arms use
Barack Obama
President **Obama** is due to sign a new nuclear pact with **Russia** this week

President **Barack Obama's administration** has unveiled a defence policy to significantly narrow the circumstances in which the **US** would use nuclear arms.

But its Nuclear Posture Review warned that countries breaking the rules would remain potential targets.

The **US** strategy document also raised concerns about a "lack of transparency" in **China**'s nuclear programme.

The review comes two days before **Mr Obama** and **his** Russian counterpart sign a landmark nuclear arms reduction pact.

The deal, agreed last month, commits **Russia** and **America** to big cuts in nuclear warheads and is to replace the 1991 Strategic Arms Reduction Treaty (Start), which expired last December.

ANALYSIS
Kim Ghattas
Kim Ghattas, **BBC** News, **Washington**

The document is carefully worded - it limits the use of nuclear weapons but carves out exceptions to use them against countries that break the rules - in other words countries like **North Korea** a
.

This new nuclear policy reflects a changing world - while nuclear weapons were useful to achieve a balance of power with countries like **Russia** or **China**, the modern threats require a more nimb
ce strategy.

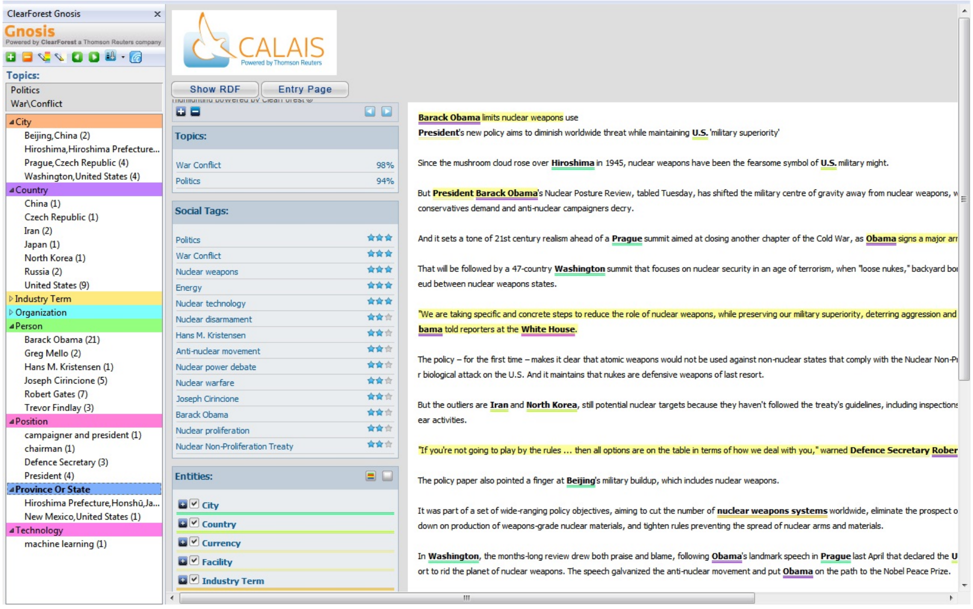
President **Obama** though is likely to be criticised by both sides- the left will say **he** should impose a blanket ban on the use of American nuclear warheads.

The right will say **he**'s undermining the credibility of **America**'s nuclear deterrent by limiting the circumstances in which it can be used.

The far-reaching Nuclear Posture Review, published on Tuesday, outlines plans for "achieving substantial further nuclear force reductions" beyond the new treaty.

Every **president** since 1991 conducts such a review - the last one took place in 2001 at the start of **George W Bush's administration**.

Figure 4.1: Enrichment of a press release with unstructured Semantic metadata with ”Document Viewer” by OpenCalais. Source: Author with OpenCalais ”Document Viewer”.


ClearForest Gnosis
Gnosis
Powered by ClearForest a Thomson Reuters company
CALAIS
Powered by Thomson Reuters
Show RDF
Entry Page

Topics:
Politics
War\Conflict

City
Beijing,China (2)
Hiroshima,Hiroshima Prefecture...
Prague,Czech Republic (4)
Washington,United States (4)
Country
China (1)
Czech Republic (1)
Iran (2)
Japan (1)
North Korea (1)
Russia (2)
United States (9)
Industry Term
Organization
Person
Barack Obama (21)
Greg Mello (2)
Hans M. Kristensen (1)
Joseph Cirincione (5)
Robert Gates (7)
Trevor Findlay (3)
Position
campaigner and president (1)
chairman (1)
Defence Secretary (3)
President (4)
Province Or State
Hiroshima Prefecture,Honshū,Ja...
New Mexico,United States (1)
Technology
machine learning (1)

Topics:
War Conflict 98%
Politics 94%

Social Tags:

Politics	★★★
War Conflict	★★★
Nuclear weapons	★★★
Energy	★★★
Nuclear technology	★★★
Nuclear disarmament	★★☆
Hans M. Kristensen	★★☆
Anti-nuclear movement	★★☆
Nuclear power debate	★★☆
Nuclear warfare	★★☆
Joseph Cirincione	★★☆
Barack Obama	★★☆
Nuclear proliferation	★★☆
Nuclear Non-Proliferation Treaty	★★☆

Entities:
☑ City
☑ Country
☑ Currency
☑ Facility
☑ Industry Term

Barack Obama limits nuclear weapons use
President's new policy aims to diminish worldwide threat while maintaining **U.S.** 'military superiority'

Since the mushroom cloud rose over **Hiroshima** in 1945, nuclear weapons have been the fearsome symbol of **U.S.** military might.

But **President** **Barack Obama**'s Nuclear Posture Review, tabled Tuesday, has shifted the military centre of gravity away from nuclear weapons, w
conservatives demand and anti-nuclear campaigners decry.

And it sets a tone of 21st century realism ahead of a **Prague** summit aimed at closing another chapter of the Cold War, as **Obama** signs a major arr

That will be followed by a 47-country **Washington** summit that focuses on nuclear security in an age of terrorism, when "loose nukes," backyard bo
eud between nuclear weapons states.

"We are taking specific and concrete steps to reduce the role of nuclear weapons, while preserving our military superiority, deterring aggression and
bama told reporters at the **White House**.

The policy – for the first time – makes it clear that atomic weapons would not be used against non-nuclear states that comply with the Nuclear Non-Pr
r biological attack on the U.S. And it maintains that nukes are defensive weapons of last resort.

But the outliers are **Iran** and **North Korea**, still potential nuclear targets because they haven't followed the treaty's guidelines, including inspections
ear activities.

"If you're not going to play by the rules ... then all options are on the table in terms of how we deal with you," warned **Defence Secretary** **Rober**

The policy paper also pointed a finger at **Beijing**'s military buildup, which includes nuclear weapons.

It was part of a set of wide-ranging policy objectives, aiming to cut the number of **nuclear weapons systems** worldwide, eliminate the prospect o
down on production of weapons-grade nuclear materials, and tighten rules preventing the spread of nuclear arms and materials.

In **Washington**, the months-long review drew both praise and blame, following **Obama**'s landmark speech in **Prague** last April that declared the **U**
ort to rid the planet of nuclear weapons. The speech galvanized the anti-nuclear movement and put **Obama** on the path to the Nobel Peace Prize.

Figure 4.2: Metadata extraction with Gnosis. Source: www.opencalais.com

4.2 IBM OmniFind Analytics

In the area of semantic search, there are already IBM's first commercial applications such as OmniFind Enterprise Edition which combined the semantic search and functional analysis or so called IBM OmniFind Analytics (see Figure 4.3). OmniFind Enterprise Edition is a software for searching the Internet, web-based CRM, blogs, wikis, forums, and file servers. OmniFind Analytics has a navigation window for the analysis of the main search results using dynamic bar charts and provides a platform for building solutions for semantic search, content analysis and visual representation in over 60 languages.

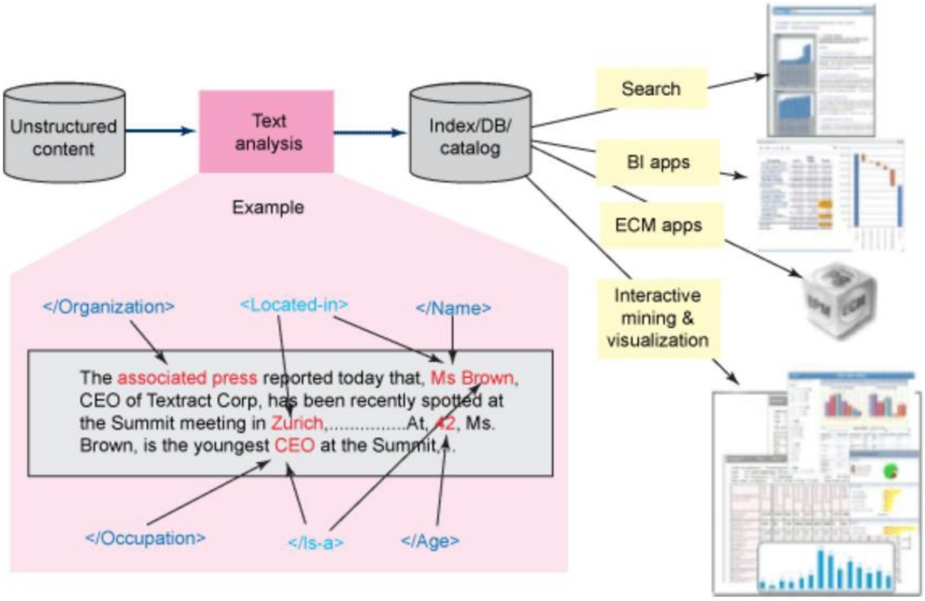
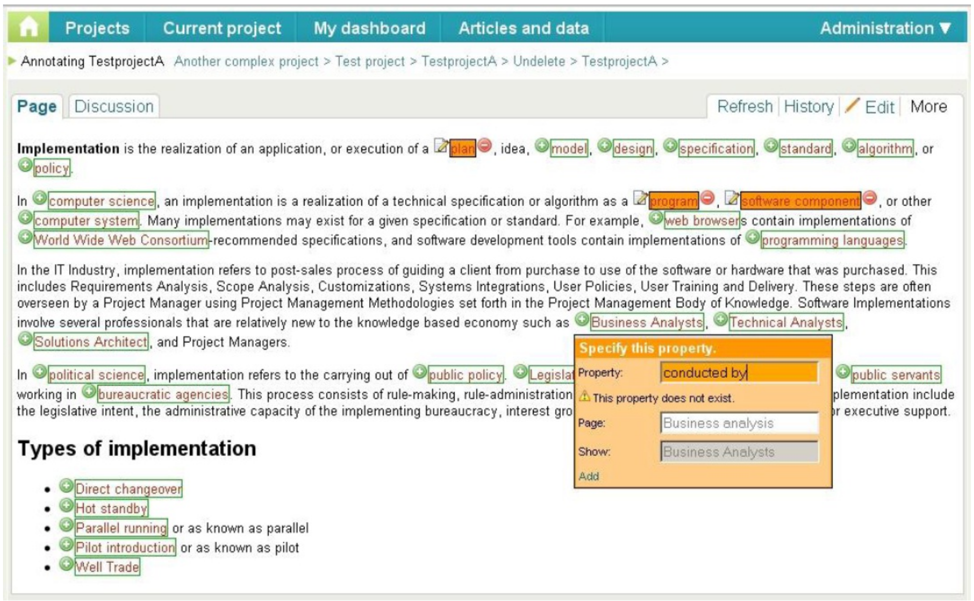

Unstructured content
Text analysis
Index/DB/ catalog
Search
BI apps
ECM apps
Interactive mining & visualization
Example
</Organization>
<Located-in>
</Name>
The associated press reported today that, Ms Brown, CEO of Textract Corp, has been recently spotted at the Summit meeting in Zurich,...............At, 42, Ms. Brown, is the youngest CEO at the Summit,.
</Occupation>
</Is-a>
</Age>

Figure 4.3: Demo version of IBM OmniFind Analytics

Semantic search and inference functions will to be integrated into master data management (Enterprise Master Data Management by IBM). IBM is currently researching on a Semantic Master Data Management (SMDM), ie to a web-based platform for master data management of customers, suppliers, accounts, products, or organizational units data.

4.3 Ontoprise SWM+

The German company Ontoprise from Karlsruhe is one of the best known providers of semantic software for organizations. Based on the Semantic Media Wiki [1] the Ontoprise offers a Semantic Media Wiki Ontoprise + (SMW +) as an organizational solution for knowledge management and project management. Compared to the open source version, many features are pre-installed, the user-friendliness of the masks and search functions were significantly improved. Thus, the annotation and search can be made without any technical knowledge through a graphical mask (see Figure 4.4).



Implementation is the realization of an application, or execution of a plan, idea, model, design, specification, standard, algorithm, or policy.

In computer science, an implementation is a realization of a technical specification or algorithm as a program, software component, or other computer system. Many implementations may exist for a given specification or standard. For example, web browsers contain implementations of World Wide Web Consortium-recommended specifications, and software development tools contain implementations of programming languages.

In the IT Industry, implementation refers to post-sales process of guiding a client from purchase to use of the software or hardware that was purchased. This includes Requirements Analysis, Scope Analysis, Customizations, Systems Integrations, User Policies, User Training and Delivery. These steps are often overseen by a Project Manager using Project Management Methodologies set forth in the Project Management Body of Knowledge. Software Implementations involve several professionals that are relatively new to the knowledge based economy such as Business Analysts, Technical Analysts, Solutions Architect, and Project Managers.

In political science, implementation refers to the carrying out of public policy. Legislat public servants working in bureaucratic agencies. This process consists of rule-making, rule-administration plementation include the legislative intent, the administrative capacity of the implementing bureaucracy, interest gro r executive support.

Property: conducted by

⚠ This property does not exist.

Page: Business analysis

Show: Business Analysts

Types of implementation

- Direct changeover
- Hot standby
- Parallel running or as known as parallel
- Pilot introduction or as known as pilot
- Well Trade

Figure 4.4: Semantic annotation of an article in SMW+, Source: http://wiki.ontoprise.com/wiki/

The search functions in addition to the full-text search include different filter masks. In this way, queries can be combined and lead to better results. When creating articles and searches the user are supported with suggestions to related articles, typo corrections, and context-sensitive request extensions. Semantic Media Wiki+ has extensive analysis and visualization capabilities that enable a quick overview of a particular topic or item (see Figure 4.5).

[1] Semantic Media Wiki , http://semantic-mediawiki.org/

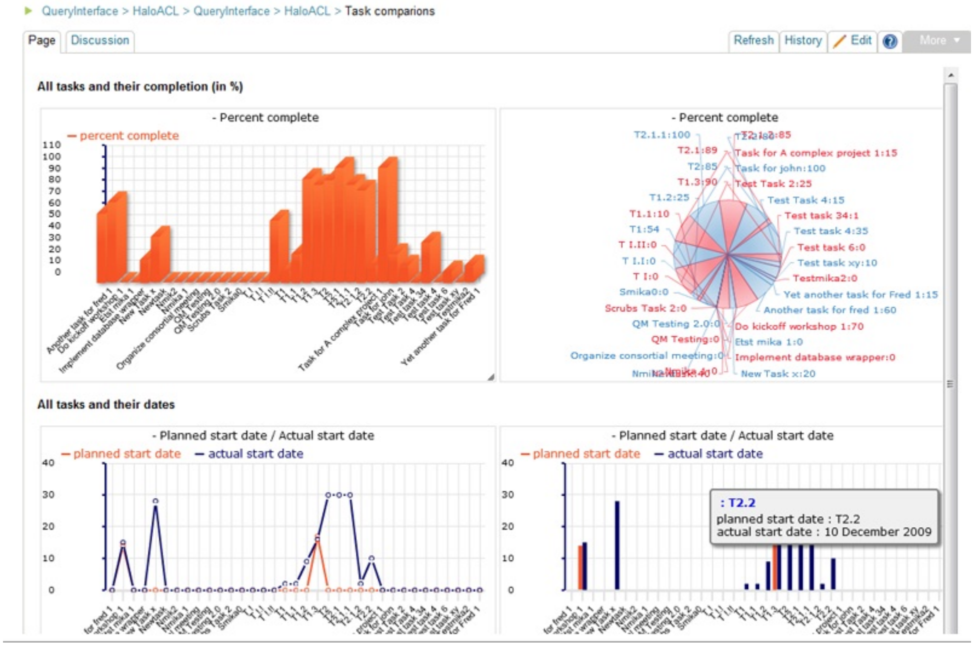


Figure 4.5: Visualization with SMW +. Source: www.ontoprise.de

4.4 Ontos Semantic API

Ontos API is a public web service which returns rich semantic metadata in standard RDF-based formats for input plain text content you submit. The technology relies on the ontology driven NLP engine OntosMiner (see Figure 4.6), which extracts information from natural language texts and creates coherent semantic metadata for the processed content. OntosMiner recognizes entities and relations between them using natural language processing techniques. Although basic types of entities (people, companies, places etc.) are pre-defined, the user can also create OWL-driven dictionaries for custom types of entities, merge entities across documents, etc. Latest versions of OntosMiner process texts in English, German and Russian.

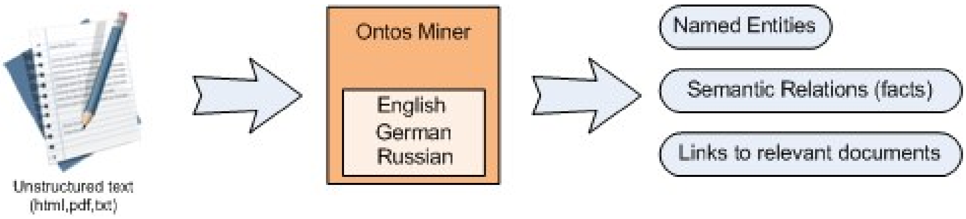
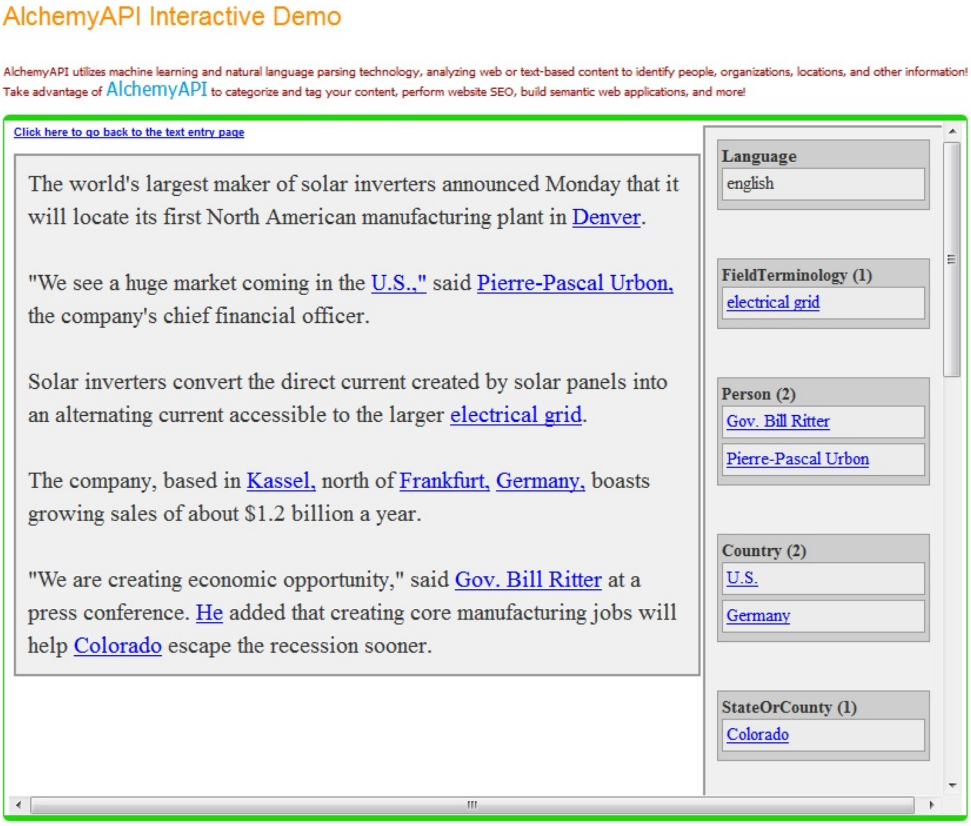


Figure 4.6: Ontos Miner

4.5 AlchemyAPI

The AlchemyAPI [2] system utilizes statistical natural language processing and machine learning algorithms to analyze your content, extracting semantic meta-data: information about people, places, companies (like OpenCalais), but also keywords describing topics. A REST API endpoint is provided and metadata is returned in XML, RDF or JSON (JavaScript Object Notation) format. To produce the keywords, statistical language processing is used, which relies on data mining techniques depending on vast volumes of training data.



AlchemyAPI Interactive Demo

AlchemyAPI utilizes machine learning and natural language parsing technology, analyzing web or text-based content to identify people, organizations, locations, and other information! Take advantage of AlchemyAPI to categorize and tag your content, perform website SEO, build semantic web applications, and more!

Click here to go back to the text entry page

The world's largest maker of solar inverters announced Monday that it will locate its first North American manufacturing plant in Denver.

"We see a huge market coming in the U.S.," said Pierre-Pascal Urbon, the company's chief financial officer.

Solar inverters convert the direct current created by solar panels into an alternating current accessible to the larger electrical grid.

The company, based in Kassel, north of Frankfurt, Germany, boasts growing sales of about $1.2 billion a year.

"We are creating economic opportunity," said Gov. Bill Ritter at a press conference. He added that creating core manufacturing jobs will help Colorado escape the recession sooner.

Language: english

FieldTerminology (1): electrical grid

Person (2): Gov. Bill Ritter; Pierre-Pascal Urbon

Country (2): U.S.; Germany

StateOrCounty (1): Colorado

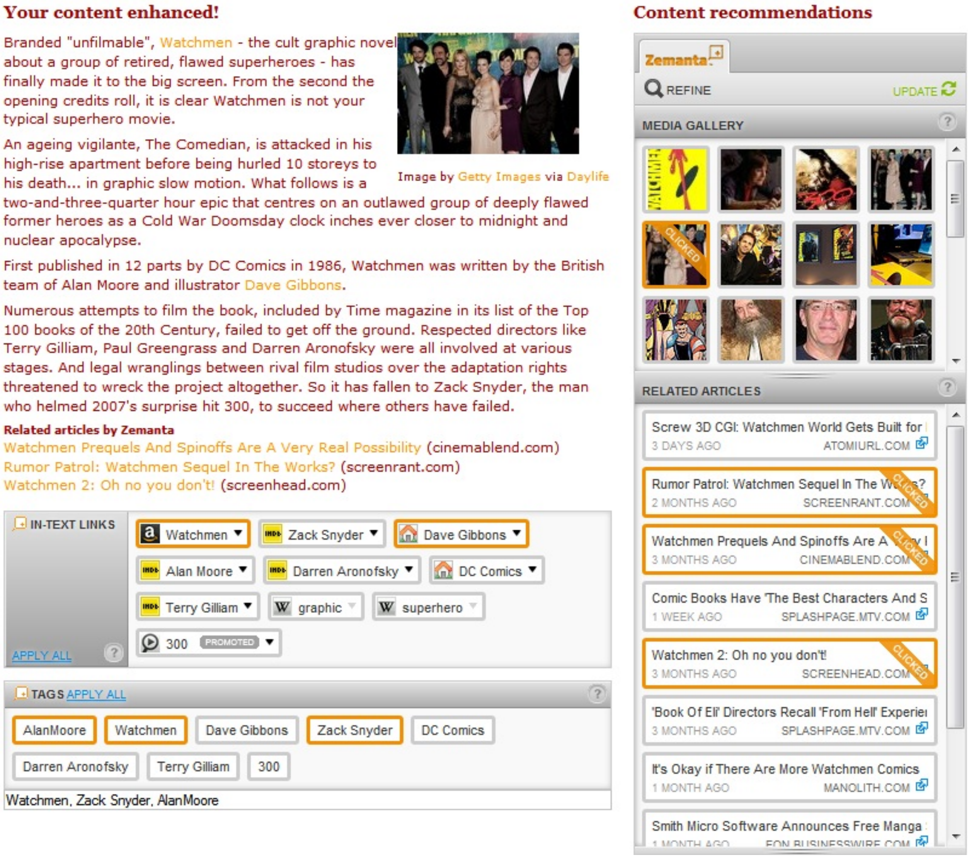
Figure 4.7: AlchemyApi Tool. Source: www.alchemyapi.com

AlchemyAPI is able to determine the language that any text or web-based content was written in. AlchemyAPI extracts keywords in eight different languages, including English, French and German.

[2] The AlchemyAPI, http://www.alchemyapi.com/

4.6 Zemanta

Zemanta offers a comprehensive range of tools for the automatic enrichment of blogs and e-mails with tags, pictures, articles from other sources, information from freely available sources such as Wikipedia, YouTube and Twitter and social networks like MySpace, Facebook, etc. Zemanta works as WordPress, Drupal or Live Writer plug-in. After installation, intuitive Zemanta delivers within seconds tags, links, photos and related items for the created text that can select and confirm the blog author (see Figure 4.8). If a ”tag” cloud is installed on the blog page, it automatically adds Zemanta new tags. Zemanta uses semantic extraction techniques to identify key issues and to contribute appropriate content from other sources. Zemanta is currently only available in English. This means that links to Wikipedia, other media and tags refer only in English-language pages. But still makes Zemanta publication deviations in corporate and private blogs much informative, efficient and visually appealing.



Your content enhanced!

Branded "unfilmable", Watchmen - the cult graphic novel about a group of retired, flawed superheroes - has finally made it to the big screen. From the second the opening credits roll, it is clear Watchmen is not your typical superhero movie.

Image by Getty Images via Daylife

An ageing vigilante, The Comedian, is attacked in his high-rise apartment before being hurled 10 storeys to his death... in graphic slow motion. What follows is a two-and-three-quarter hour epic that centres on an outlawed group of deeply flawed former heroes as a Cold War Doomsday clock inches ever closer to midnight and nuclear apocalypse.

First published in 12 parts by DC Comics in 1986, Watchmen was written by the British team of Alan Moore and illustrator Dave Gibbons.

Numerous attempts to film the book, included by Time magazine in its list of the Top 100 books of the 20th Century, failed to get off the ground. Respected directors like Terry Gilliam, Paul Greengrass and Darren Aronofsky were all involved at various stages. And legal wranglings between rival film studios over the adaptation rights threatened to wreck the project altogether. So it has fallen to Zack Snyder, the man who helmed 2007's surprise hit 300, to succeed where others have failed.

Related articles by Zemanta

Watchmen Prequels And Spinoffs Are A Very Real Possibility (cinemablend.com)
Rumor Patrol: Watchmen Sequel In The Works? (screenrant.com)
Watchmen 2: Oh no you don't! (screenhead.com)

Watchmen, Zack Snyder, AlanMoore

Content recommendations

Screw 3D CGI: Watchmen World Gets Built for
3 DAYS AGO ATOMIURL.COM

Rumor Patrol: Watchmen Sequel In The W ?
2 MONTHS AGO SCREENRANT.COM

Watchmen Prequels And Spinoffs Are A y I
3 MONTHS AGO CINEMABLEND.COM

Comic Books Have 'The Best Characters And S
1 WEEK AGO SPLASHPAGE.MTV.COM

Watchmen 2: Oh no you don't!
3 MONTHS AGO SCREENHEAD.COM

'Book Of Eli' Directors Recall 'From Hell' Experie
3 MONTHS AGO SPLASHPAGE.MTV.COM

It's Okay if There Are More Watchmen Comics
1 MONTH AGO MANOLITH.COM

Smith Micro Software Announces Free Manga
1 MONTH AGO FON BUSINESSWIRE.COM

Figure 4.8: Creating a blog post with Zemanta, Source: www.zemanta.com

Professional users such as publishers and media organizations also have the opportunity to upload their own archives and set up their own recommendation systems to rely not only on publicly available data and automatically generated keywords.

In April 2009, Zemanta was able to access over 100 million annotated objects, including message content and selected blogs. Building on semantically annotated databases (RDF) Zemanta uses ”Natural Language Processing” extraction procedures in order to identify the most relevant content using semantic analysis. The underlying semantic technology then subscribed to the feeds at different keywords of the bookmarking services and other verified sources and delivers the content via a user interface - ”On-Demand”.

4.7 Summary

This chapter presents shortly the currently existing applications for semantic enrichment. It shows aspects like semantic knowledge representation of the content on the World Wide Web, semantic search, content analysis and visual representation. These already existing applications have reflected to the achievement of this thesis's work.

Chapter 5

SemTex Application

> *" Software suppliers are trying to make their software packages more "user-friendly... Their best approach so far has been to take all the old brochures and stamp the words "user-friendly" on the cover."*
>
> [Bill Gates.]

This chapter introduces the SemTex (Semantic Text Analysis and Interlinking) Web application - A novel approach of accessing data on the Web. Then, it presents the global concept proposed by this thesis following with the scenarios using Linked Data, the system architectural design, use cases and implementation as a part of a research project.

5.1 Introduction

The current business model of commercial content providers, e.g. online newspapers, is comparatively transparent: Content providers aim to produce useful contents, publish these contents on their portals and indirectly monetize them by serving advertisements. Hence, revenues of online providers largely depend on the number of users consuming online content (their reach) and indirectly also on their session length.

The two most common advertising pricing models are the cost per impression (CPI) and the cost per click (CPC) model. Advertisements are either served directly by the content provider or by third party services. For a more detailed discussion on business models, cf. e.g. [Lyons et al., 2009] or [Osterwalder and Pigneur, 2002]. Valuable online content, which is suitable for monetarization through advertisement, is in practice created by a specialist, the online editor.

An online editor usually investigates upcoming topics, aggregates (online) content from various sources, including user generated content taken from blogs, or professional content from press agencies, and merges all these small junks together shaping a fascinating story, capable of drawing the attention of the user. Needless to say, online editors require a good nose for how to create such content, which is preferably consumed by people on the Web.

Unlike any other role in the online content industry, professional online editors depend and rely more on Web technology. Research has shown that people on the web are rather scanning content than reading everything in detail which is contradictory to readers of classical newspapers [Holmqvist et al., 2003].

Though, very little is known about the needs of people regarding nature, structure and presentation of online content, and we may just imagine, what differs good online content from bad. In practice, the amount of revenue generated from advertisements may serve as indicator, determining quality and appropriateness of online content.

Human resources are always scarce, which implies that they have to be utilized effeciently and effectively. From talks with commercial online content providers we learned that professional online editors spend most of their time investigating interesting and suitable material, which may enrich their editorial content to become more fascinating and valuable to their audience and/or enables them to produce editorial content quicker, enriched with multimedia objects and hyperlinks to advanced material [Halb et al., 2010].

The latter is, where third party content will play an important role. Accurately considering these aspects may increase the attention of existing content consumers, keeping them on site for a longer period as well as attracting new consumers, positively affecting the amount of revenues gained from ads embedded in the content. Anyway, to prove these hypotheses is not the goal of this thesis.

The research problem may be outlined as follows: Professional online editors are facing at least two business challenges:

- They need to be very efficient in their core business, i.e. developing their editorial content and getting it published instantly on the Web.
- They have to produce up to date (multimedia) content, particularly capable of drawing the attention of humans.

We investigated methods for text analysis in chapter 2 and current Web technology in chapter 3, foremost Linked Data, to support online editors in creating professional online ontent.

Motivated by the challenges mentioned above, a SemTex prototype has been developed that is intended to support editors at online content providers but can also be used as a general purpose tool in other domains. It enriches editorial content with further information from Linked Data sources and also generates a Linked Data version of the editorial content. A distinguishing feature of the tool is that it supports multilingual content which poses additional challenges in the context of the Web of Data where English is still the predominant language. Currently the SemTex prototype is being tested at a cooperating content provider and will be made public in the near future.

SemTex project is an ongoing research project supported by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit).

The next sections explain in details the whole concept, architecture, uses cases, and implemented modules of SemTex prototype using of Linked Data for business and transforming them to specific situation.

5.2 Scenarios using Linked Data as a technology-oriented solution

Linked Data (see section 3.5.1) as subtopic of Semantic Web based on four simple rules (introduced by Tim Berners-Lee) is capable of dealing with both introduced challenges from the previous section. In a nutshell, the main objective of the Linked Data community is to first generate a semantically enriched Data (or Web of Data) and from the results generated others can easily build intelligent applications on behalf of this data.

Such a strategy is supposed to be very useful in solving problems in at least three different business scenarios [Halb et al., 2010]:

1. Enterprises may use Linked Data to interlink/enrich their own content, increasing its accessibility for humans and machines.
2. Enterprises may use Linked Data to integrate third party content into their own portals.
3. Enterprises may use Linked Data to prepare their own content for third party adoption, enhancing its reusability and visibility.

These three general scenarios will benefit especially to commercial content providers, as they are dealing with large amounts of data usually stored in external storages (data silos), finding themselves against at least one of the following scenarios:

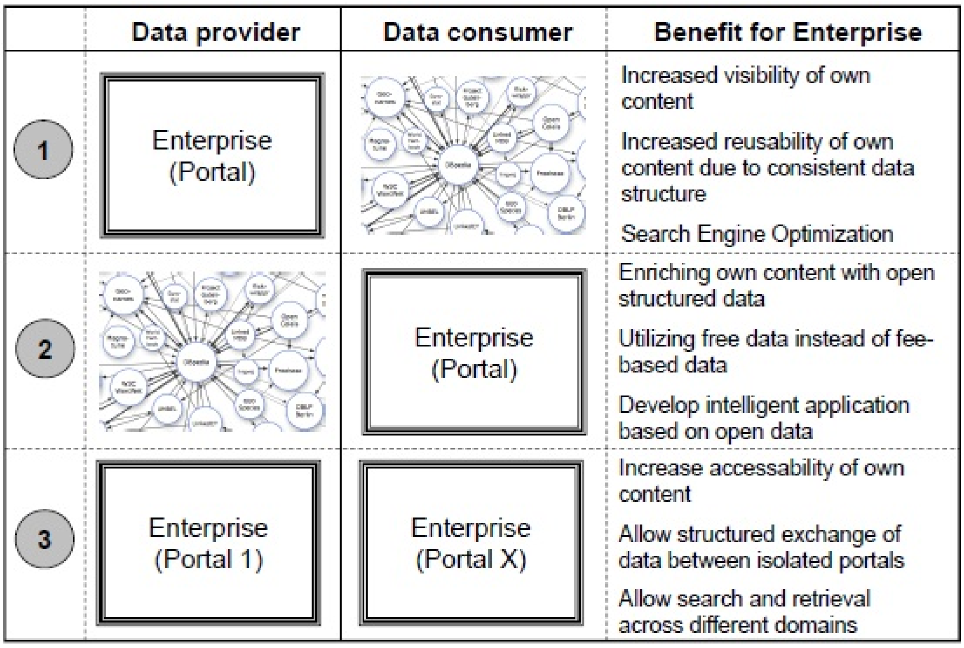
1. They may operate more than one content portal, facing the need to better integrate and interlink their data to achieve better accessibility, i.e. to allow enhanced search and retrieval across portals.
2. They may want to integrate third party data to enrich their own editorial content with open structured data, choose open - instead of licensed - content to reduce costs, and develop valuable intelligent applications based on open data.
3. They may provide their own content to be used by third parties, thereby achieving a significant increase in visibility and reach, raising the general reusability of their own content, and leading to third party adoption coming along with search-engine related benefits.

Figure 5.1 shows the benefits for online content providers that may use Linked (Open) Data. Using Linked Data may definitively result in a paradigm shift for the professional business: Online editors may benefit much from Linked Data as this new technology will enable them to perform better and to create more valuable content. On the one hand, Linked Data will enable them to integrate their own or third party data to generate appropriate and up-to-date content, and on the other hand has already resulted a plethora of different data beginning from geographical to statistical, ready for integration.

Linked Data may have more value to the human end user, than ordinary data. Figure 5.2 explains the Linked Data Value Chain, a lightweight model, conceptualizing the business perspective of Linked Data and making the value adding process to the data transparent.

The Linked Data Value chain [Latif et al., 2009] is built on three different concepts: Participating Entities, Linked Data Roles, and Types of Data. Participating entities are persons, enterprises, associations or research institutes and can act at least one of the following roles:

- Raw Data Providers provide all kinds of data except RDF format.
- Linked Data Providers provide any kind of data in a Linked Data format by tranforming the data consumed from Raw Data Providers.



	Data provider	Data consumer	Benefit for Enterprise
1	Enterprise (Portal)		Increased visibility of own content Increased reusability of own content due to consistent data structure Search Engine Optimization
2		Enterprise (Portal)	Enriching own content with open structured data Utilizing free data instead of fee-based data Develop intelligent application based on open data
3	Enterprise (Portal 1)	Enterprise (Portal X)	Increase accessability of own content Allow structured exchange of data between isolated portals Allow search and retrieval across different domains

Figure 5.1: Three scenarios for the use of Linked Data. [Halb et al., 2010]

- Linked Data Application Providers provide Linked Data Applications. They consume Linked Data provided by Linked Data Providers, process it within their applications and transform it into Human-Readable-Data.
- End users are humans who (like to) consume Human-Readable-Data, which is a human-readable presentation of Linked Data provided by Linked Data Application Providers.

Types of Data of Linked Data Value Chain includes:

- Raw Data is any kind of structured/unstructured data.
- Linked Data is Raw Data converted in RDF format interlinked with other RDF data
- Human-Readable-Data is any kind of data readable by humans.

Regarding to the introduced Linked Data Value Chain, most of online content providers currently provide Raw Data acting as Raw Data Providers except BBC [Kobilarov et al., 2009] and the New York Times [Larson and Sandhaus, 2009]. Linked Data technology allows enterprises to act as Linked Data Providers, providing Linked Data for third parties, and to act as Linked Data Application providers, consuming data from third parties and providing more valuable Human-Readable-Data.

Based on the business challenges of online editors and the concepts of Linked Data Value Chain, the first prototype of SemTex application is developed. The next sections explain the architecture and implementation of SemTex application.

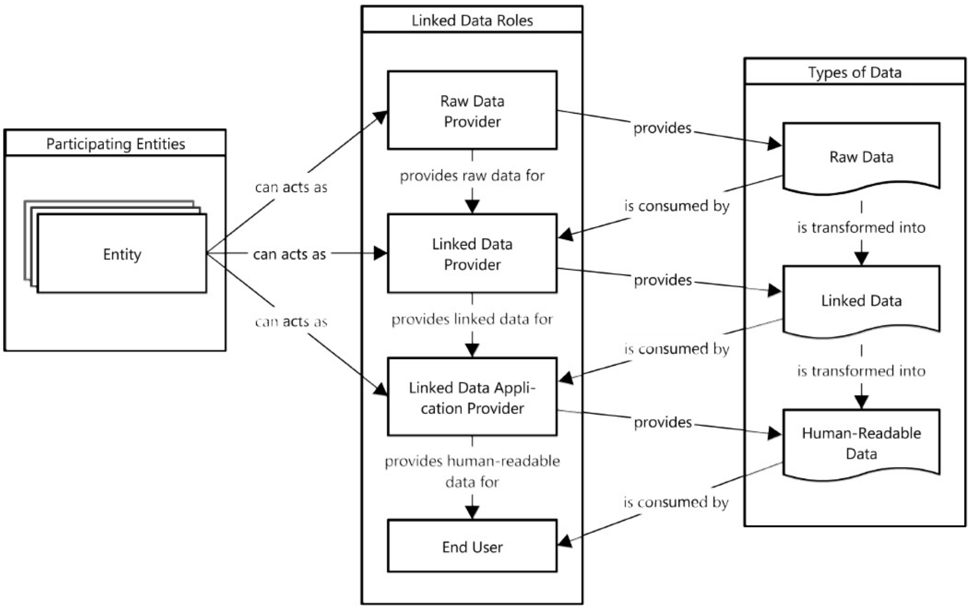


Figure 5.2: The Linked Data Value Chain. [Latif et al., 2009]

5.3 System Architecture

> *" A software architecture is a description of the subsystems and components of a software system and the relationships between them. Subsystems and components are typically specific in different views to show the relevant functional and nonfunctional properties of a software system. The software architecture of a system is an artifact. It is the result of the software design activity. Software design is the activity performed by a software developer that results in the software architecture of a system. It is concerned with specifying the components of a software system and the relationships between them within given functional and non-functional properties. "*
>
> [Buschmann et al., 1996]

Before starting into details with the implementation of SemTex Web application, let us understand technical environment, general concepts of aims introduced by this thesis as a part of a research project. The following will be discussed about the architectural design that deals with subsystems or modules of a SemTex Web application, their relationships, and interactions between them.

5.3.1 Overview of system design

To decide which software architecture should be used for constructing various Semantic Web applications or Linked Data-Driven Web applications is not an easy task. Most Semantic Web applications are constructed using the same fundamental principles, similar components, and variations of a basic architecture. In a linked data-driven Web application conceptually one will be able to identify the following components [Hausenblas, 2009]:

- A local RDF store , able to store results and act as a permanent storage device. We note that an RDF store such as ARC2 or Virtuoso is not a strict requirement, though often it makes sense to manage the RDF data in a native environment
- Some logic (a controller) and UI components implementing the business logic, the User Interface (UI) and the interaction parts of the application. These components are not specific to linked data-driven Web applications, however typically required and found in the wild.
- A data integration modul, focusing on fetching linked data from the Web of Data, either directly from the LOD cloud or via Semantic Indexer such as Sindice or Falcons.
- A republishing component that eventually exposes parts of the application's (interlinked) data on the Web of Data. It is a good practice to republish the application's data, hence providing again input to the LOD cloud.

The above mentioned components do not dramatically differ from what typical XAMP-based Web applications [1] look like. The two main differences can be seen in the data integration and the republishing module. While the latter may well be a sub-module of the logic and/or UI module, the former characterised a linked data-driven Web application. In 5.3 the modules and interaction between external entities are depicted.

Based on the concept of linked data-driven Web application from Hausenblas [2009], the SemTex project aims to develop a semantic and largely automatic networking of media information, taking into account of the user's specific context of production. In the figure 5.4 whole proposed system design of the SemTex system is depicted. SemTex system consist of four components:

[1] XAMP, http://en.wikipedia.org/wiki/xamp

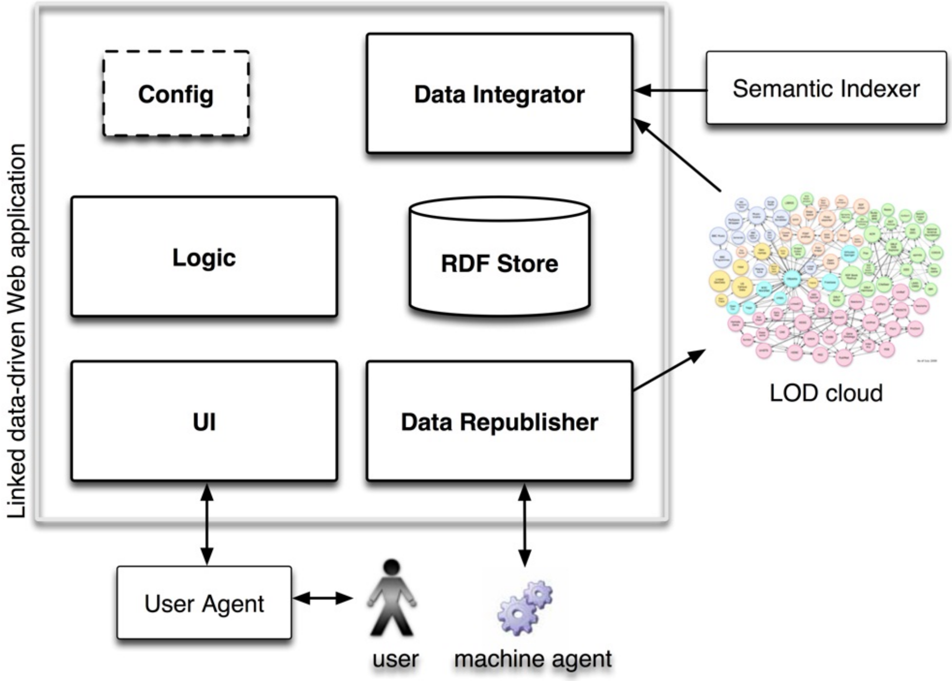


Figure 5.3: Concept of a linked data-driven Web application [Hausenblas, 2009]

- SemTex UI
- Term extraction and classification,
- Linked Data consumation and interlinking ,
- Data Publication.

We note that the SemTex system has no separated Data Integrator. The Data integrator is integrated in Linked Data consumation and interlinking component. Unfortunately , some of Linked datasets such as GeoNames, DBPedia can not be always accessed through online SPARQL Endpoints. But the Linked Datasets provide another way of accesing its data via datasets dumps. Therefore, as an optional solution of fetchind data from Linked Data, the SemTex prototype has imported these data into a local RDF store such as Virtuoso or TDB Jena.

The *SemTex UI* component is dedicated for online editors to write articles including multimedia information. This *SemTexUI* component communicates with other components such as: *Term extraction and classification*, *Linked Data consumation and interlinking* and *Data Publication* component, as it depicted in 5.4. The communications between *SemTex UI* and *Term extraction and classification* module is done using the Simple Object Access Protocol(SOAP). The first input of SemTex UI module recieves results from the Term extraction and classification module which contains the interesting terms extracted and classified from the edited text. The second input of SemTex UI cames from *Linked Data consumation and interlinking* module which uses the extracted terms as an input from Term and classification module. The *Linked Data consumation and interlinking* retrieves additional information from Linked Data about terms extracted. The last component is the *Data Publication* component. It recieves the final user article enriched with further information from the SemTex UI component and publishes considering the Linked Data principles and interlinking it with other existing datasets.

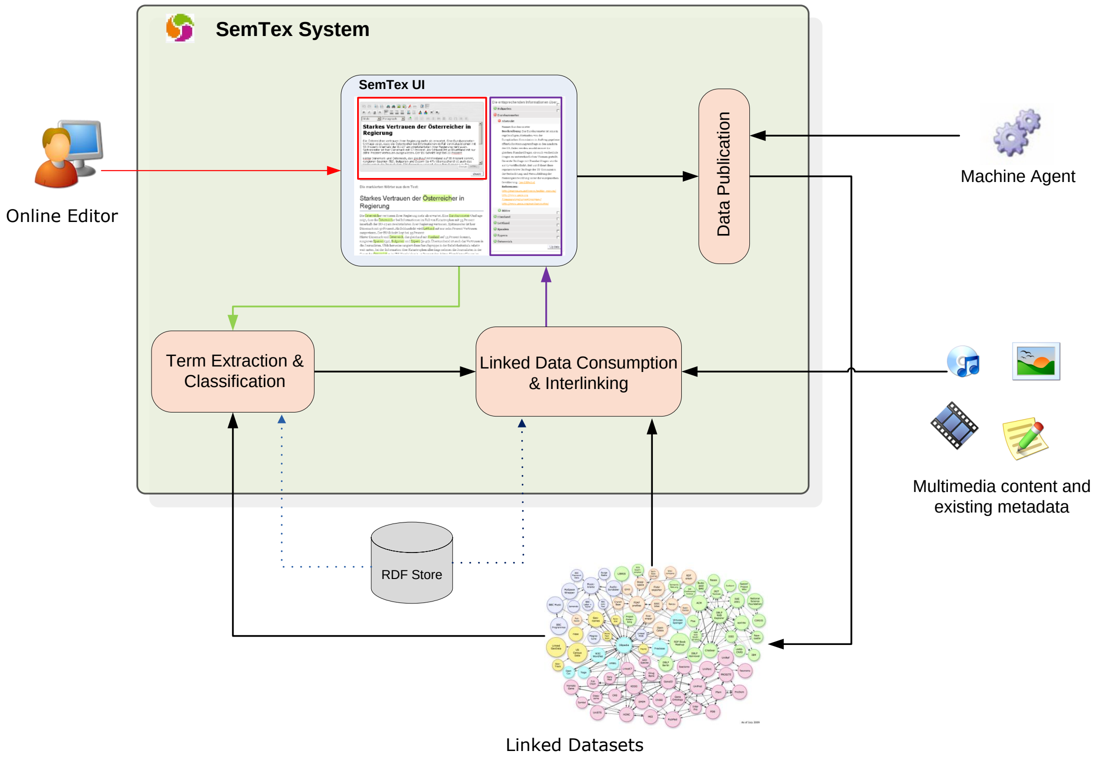

SemTex System
SemTex UI
Online Editor
Data Publication
Machine Agent
Term Extraction & Classification
Linked Data Consumption & Interlinking
Multimedia content and existing metadata
RDF Store
Linked Datasets

Figure 5.4: SemTex Architecture

5.3.2 System components

As it depicted in 5.4 , the SemTex System consists of four distinct components that work together as a unified solution that can be integrated into an existing content management system as a service or can be used as a stand-alone web application. The term extraction and classification module recognizes interesting and relevant terms which act as an input for the Linked Data consumption and interlinking module where additional information from Linked Data sources is collected. The Data Publication module makes the discovered information available as Linked Data to the public.

5.3.2.1 SemTex UI

The SemTex UI is the main component used to create user's content. It uses other components to analyse the user's content and to enrich the user's content with further information from the Web. In order to create the user's content, this component has integrated an online YUI [2] Rich Text Editor (see 5.5.5.7). It allows online user to edit rich text within web browsers. The detailed explanation of SemTex UI will be followed by section 5.7.5

5.3.2.2 Term extraction and classification

The term extraction and classification component that has been developed is targeted at German-speaking content. Even though there exist several term extraction solutions for the English language (which can also be plugged in and used in the prototype) there is still a lack of well-performing, generally available

[2] YUI, http://developer.yahoo.com/yui/

tools for German. Through its modular structure the Semtex prototype can support term extraction solutions for any language that are available from third-party providers. The developed methods of this module will be introduced in section 5.7.2.

5.3.2.3 Linked Data consumption and interlinking

The Linked Data consumption and interlinking component retrieves supplementary information about the identified terms and creates interlinks to Linked Data sources. The general work flow of the component starts with a search for potentially relevant information in external data sources. In the first phase this is done via a query to DBpedia and GeoNames, two datasets that already contain a lot of general knowledge and geographic information. Even though the use of further datasets is easily possible, this first step aims at identifying relevant concepts for the found terms where these two Linked Data hubs already provide sufficient information for general news articles. The query is based on the extracted term and takes the original document's language into account as well as textual similarity measures.

Once the concept has been identified it is possible to retrieve further information from different data sources. The information gained from Linked Data sources can also be extremely useful for queries to other repositories containing for instance user contributed multimedia content. It is possible to find more appropriate matches as the query can be enhanced with more metadata for finding relevant images and videos. Especially for concepts related to geographical locations it is possible to supply coordinates that have been retrieved from DBpedia or GeoNames in the query when searching for images on Flickr or videos on Youtube.

5.3.2.4 Data Publication

The final article contains related information from Linked Data sources as well as image and video content from Flickr and Youtube. It also includes an RDF representation that is embedded in the webpage via RDFa, a practice for building linked data for both humans and machines as described in Halb et al. [2008]. The recently added support of RDFa content by important global search engine providers such as Yahoo![3] and Google[4] underlines the industry's appreciation of this approach. Further an RDF/XML version is also provided.

[3] Yahoo, http://www.yahoo.com

[4] Google, http://www.google.com

5.4 Use Cases and Requirments

The primary use case of the developed prototype is to support the editorial process at online content providers. In order to meet the needs of the industry we collaborated closely with one of the major content providers in Austria.

The SemTex prototype can be integrated in a provider's content management system and while an editor creates an article, the text is analyzed for interesting terms and the tool automatically suggests further information from the Web of Data. This includes data from various Linked Data sources such as textual content, links, images, or video.

The online editor instantaneously receives additional information about the article she is writing and can then decide which external content should be included in the article. This manual decision step has been introduced following feedback from editors and content providers as they prefer to have more control over the published content.

It further allows to improve content suggestions based on previous selections and preferences. As an added benefit the final article is enriched with more exciting content and provides the reader with further information without having to leave the provider's pages. This increases the attractiveness of the content provider with further potential positive effects on visit durations, returning users, as well as page and add impressions leading to higher add revenues for the content provider. The figure 5.5 depicts an use case diagram of SemTex prototype.

The requirments that fulfill these use cases are based on the objectives from the chapter 1. Each Objective has its own requirments as following:

Objective 1: Term extraction and classification has the following requirments:

1. Term recognition/extraction
2. Term classification

Objective 2: Linked Data Consumation and Interlinking has the following requirments:

1. Data identification
2. Data retrieval/consumation
3. Storing Data
4. Interlinking of data

Objective 3: Data Publication has the following requirments:

1. Creating XHTML Page with embedded RDFa
2. Creating RDF/XML

Objective 4: SemTex UI has the following requirments:

1. User Input Text Editor
2. Linked Data Visualisation

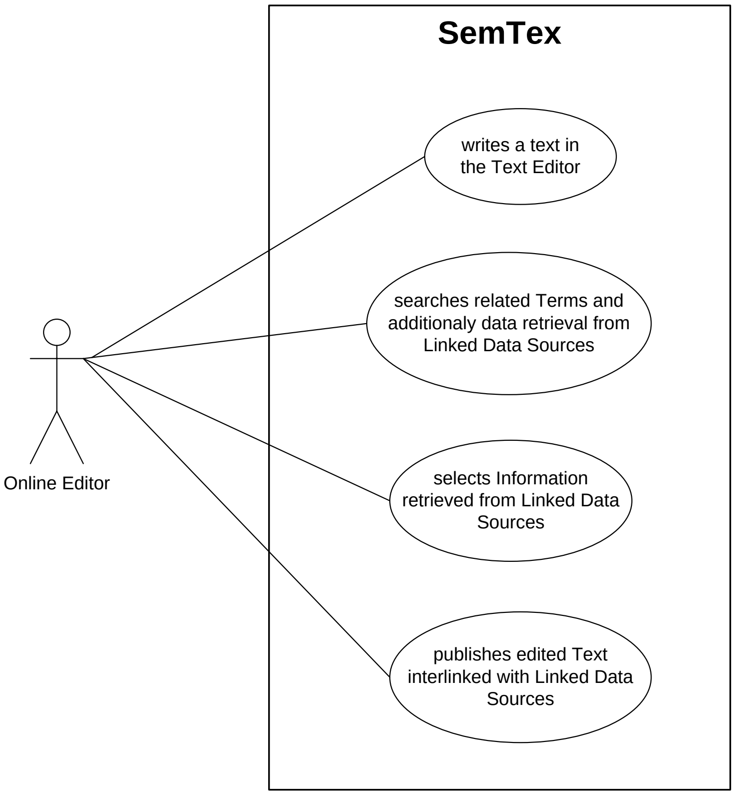

SemTex
writes a text in the Text Editor
searches related Terms and additionaly data retrieval from Linked Data Sources
selects Information retrieved from Linked Data Sources
publishes edited Text interlinked with Linked Data Sources
Online Editor

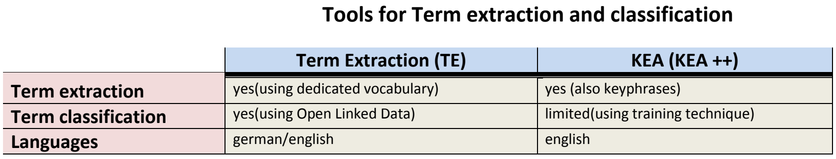
Figure 5.5: Use Case Diagram of the SemTex Prototype

5.5 Technology Selections and Evaluation

Selecting the correct technologies is the most important and complex process to fulfill the requirments of this work which are mentioned in section 5.4. Before selecting a technology, we must be sure that the technology contributes to the success of one of our objectives. Therefore, the technology must be checked from both development and architecture perspective. That means if the chosen technology will be integrated with our project.

5.5.1 Evaluation of tools and techniques for term extraction classification

Term extraction is used to extract relevant terms from text. There are many tools for term extraction which some of them do not suit to the SemTex application. Therefore, only the Term Extraction tool and KEA have been evaluated as depicted in figure 5.6.



Tools for Term extraction and classification

	Term Extraction (TE)	KEA (KEA ++)
Term extraction	yes(using dedicated vocabulary)	yes (also keyphrases)
Term classification	yes(using Open Linked Data)	limited(using training technique)
Languages	german/english	english

Figure 5.6: Feature comparison between TE and KEA

TE(Term Extraction) tool from 5.7.2 is developed by Joanneum Research Group[5] that is targeted at german-language. Even though there exists several term extraction solutions in english language, there is still a lack of well-performing, generally available tools for german. Candidate terms were filtered using all possible morphological formations from a controlled dictionary. As a controlled dictionary a full word list of ”Wahrig” [6] german dictionary has been used. This term extraction technique do not require any training data; in other words they are unsupervised.

Kea, the Keyphrase Extraction Algorithm is based on similar principles but using a different extraction technique. The textual sequences are defined by orthographical boundaries such as punctuation marks, numbers, and newlines. During the filtering phase, two features for each candidate are processed: the TF×IDF measure and the position of first occurrence. A Naïve Bayes classifier analyzes training data and creates two sets of weights: for candidates matching manually assigned keyphrases and for all other candidates. In the filtering stage, the overall probability of each candidate being a keyphrase is calculated based on these weights. The candidates are ranked according to their probabilities. This method require a lot of training and testing to achieve the correctness of keyphrases; in other words they are supervised.

Text Extraction tool employ more accurate term generation techniques than KEA. While the majority of machine learning approaches like KEA simply extract word n-grams, heuristic shows that Term Extraction tool compensates the lack of training data by complex analysis using shallow parsing by [Barker and Cornacchia, 2000], morphological conflation by [Paice and Black, 2003] and reference corpora.

Term Extraction lies outside of the scope of this thesis because it was already implemented. Term Extraction modul has been integrated in SemTex project using SOAP commands as explaned in section 5.5.5.3. The goal for this part of master thesis was to identify and classify the extracted terms using Linked Open Data. In other words to filter the black list of extracted terms. Wikipedia, an online encyclopia as a popular source of linguistic knowledge in many languages has ben chosen to fulfil this

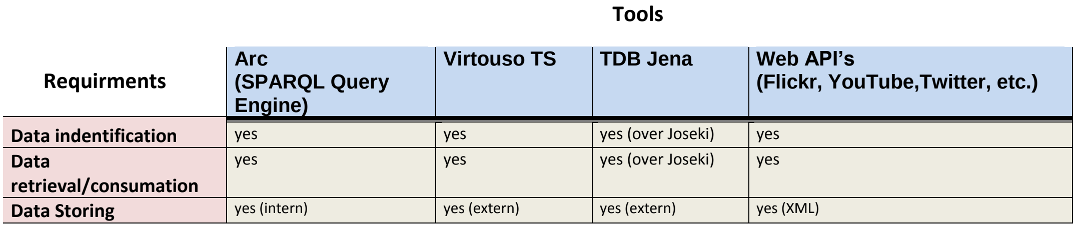
[5] Joanneum Research, http://www.joanneum.at

[6] Wahrig Dictionary, http://www.babylon.com/dictionary/wahrig/german/

goal where terms are equated. Individual terms correspond to common language concepts and named entities, disambiguation list meanings of ambiguous terms, and links between concepts that indicate semantic relatedness.

5.5.2 Tools selections and evaluations for Linked Data Consumation and Interlinking

Consuming and interlinking of semantic data on the web is still a challinging part because there are not enough linked data on the web. So that there are just a few tools which I have compared and evalated as are depicted in figure 5.7. The figure shows the requirments from objective 2 (see section 5.4) and the tools which lead to their completeness. The evaluation is done based on the success of requirment. The most used technology is ARC2 (as explained in the the section 5.5.5.1). The ARC2 has the ability to access Linked Data via HTTP URIs using queries on remote SPARQL endpoints that are made available by Linked Data consolidators(such as Virtouso instance hosting Linked Open Data [7]). Using remote SPARQL endpoint from DBPedia or Geonames help us to identify/disambiguate terms and retrieve additional information about the extracted terms, but also to classify the terms.



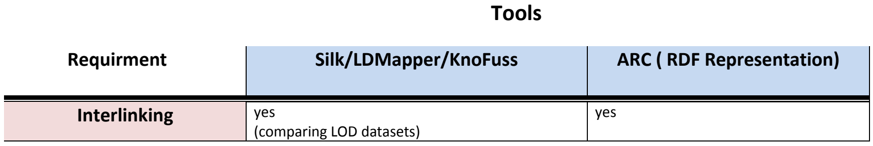
	Tools			
Requirments	Arc (SPARQL Query Engine)	Virtouso TS	TDB Jena	Web API's (Flickr, YouTube,Twitter, etc.)
Data indentification	yes	yes	yes (over Joseki)	yes
Data retrieval/consumation	yes	yes	yes (over Joseki)	yes
Data Storing	yes (intern)	yes (extern)	yes (extern)	yes (XML)

Figure 5.7: Evaluation of Linked Open data tools towards completeness of requirements

The information gained from Linked Data are extremely useful for queries to other repositories containing user contributed multimedia content. Especially for terms related to geografical locations it is possible to supply coordinates that has been retrieved from DBPedia or Geonames in the query when searching for images and videos. Unfortunatly a lot of repositories have data silo that means data are not semantically linked. Therefore, the only way to connect to such repositories like YouTube or Flickr for retrieving videos or images are used Web API's such as RESTful-Web-APIs. Zend GData from 5.5.5.2 is strongly used in this case to retrieve videos from YouTube.

Interlinking of consumed data is strongly dependent on ARC2. Before publishing the document with enriched data from the web, we manually embedd RDF attributes (see section 5.7.4) based on Linked Data principle. This document that is a XMTML+RDFa document is parsed using ARC2. After parsing of XMTML+RDFa document a RDF/XML representation is generated which is a practice for building linked data for both humans and machines. The approaches about link discovery that have been proposed for Linked Data sources such as SILK, KnoFuss and LD-Mapper are analysed from their perspectives but not used for interlinking. These approaches rely on further information about a resource that can be compared in two different datasets, for example there needs to be some overlap between sources and target datasets which is not the case in our scenario where the source contains no more information about concepts than its label.

[7] LOD Virtouso, http://lod.openlinksw.com



	Tools	
Requirment	Silk/LDMapper/KnoFuss	ARC (RDF Representation)
Interlinking	yes (comparing LOD datasets)	yes

Figure 5.8: Evaluation of tools for interlinking

5.5.3 Tools for data publication

The data publication includes the tools to publish information conform to Linked Data principles using formats that allows machines to access and re-use the information. The Figure 5.9 introduces the tools imposed by SemTex application on selection and evaluation for data publication.

These tools are devided on three main different aspects:

1. **In-Browser tools** - allow detecting and parsing of RDFa markups embedded in XHTML. Fuzz[8] is implemented as a Firefox Add-on that performas actions on the semantic data using a custom UI specific to the data to the person browsing. On the other side RDFa Highlight[9] bookmarklet is javascript implemtation that enables you to examine and validate the RDFa markup. Since the Rdf attributes are manually added in SemTex as described in section 5.7.4, the best way to speed up the testing and validating of the Rdfa was Rdfa bookmarklet. Fuzz is more an advanced tool to automatically detect exact information from the website based on the searched term and takes specific actions on demand.

2. **Content Management System tools** - that allow us to integrate the RDF attributes in SemTex pages by an site operator is not an easy task. It requeries a lot of interest in learning the details of RDF and description logics. Therefore, it has to be provided by the software or plugins. The process of integrating RDF attributtes has to be simple and straightforward. Drupal facilitates the creation of web sites by integrating Rdfa (Corlosquet et al.). Drupal CCK [10] content model consists of content types, fields, and nodes that instantiate the types, which is good way of representing it in RDF. Mapping the site data model to existing ontologies, the site operator first imports the ontology. Then she adds for every content type and field with a property or class that can be chosen from the ontology it should be mapped to. Wordpress is another open source CMS and publishing platform. There are already a few RDFa editors plugins Khalili and Auer [2011] for Wordpress such as RADiFy,WYMeditor and RDFaCE. But still some of them have drawbacks: no editing option to edit the created triples, difficulties to annotate a part of text within a block or intead of "rel" attribute uses the "property" attribute for all predicates. SemTex application is not integrated into any of these CMS's. Drupal will be an solution for Semtex application because is more RDF-based and it shows easy way to add various meta descriptions for content.

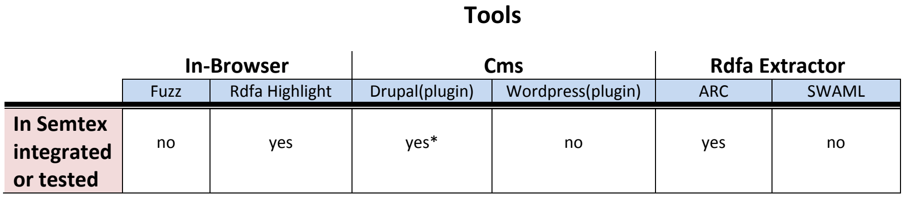
3. **RDFa Extractors** - allow to extract set of RDF triples from an RDFa annotated XHTML. Unfortunately, there are no general tool to extract any RDF triples from an XHTML+RDFa page. There are not many tools that can used with SemTex. Most of them are bookmarklets, plugins, online testers

[8] Fuzz, http://rdfa.digitalbazaar.com/fuzz/trac/

[9] RDFa Highlight, http://www.w3.org/2001/sw/BestPractices/HTML/rdfa-bookmarklet/

[10] Drupal CCk, http://drupal.org/project/cck

with no API SWAML is a mail list exporter. It extracts a collection of email messages stored in a mailbox and generates a RDF triples. SWAML is written in Python using SIOC ontolgy. ARC2 offer more than SWAML. Using ARC2 methods as discribed in 5.5.5.1 every XHTML-RDFA page can be parsed using dedicated parsers for a specific serialization. ARC supports 4 serialization formats: RDF/XML, RDF/JSON, Turtle and N-Triple.



Tools

	In-Browser		Cms		Rdfa Extractor	
	Fuzz	Rdfa Highlight	Drupal(plugin)	Wordpress(plugin)	ARC	SWAML
In Semtex integrated or tested	no	yes	yes*	no	yes	no

Figure 5.9: Tools for data publication

5.5.4 Frameworks selections and evaluation of SemTex UI

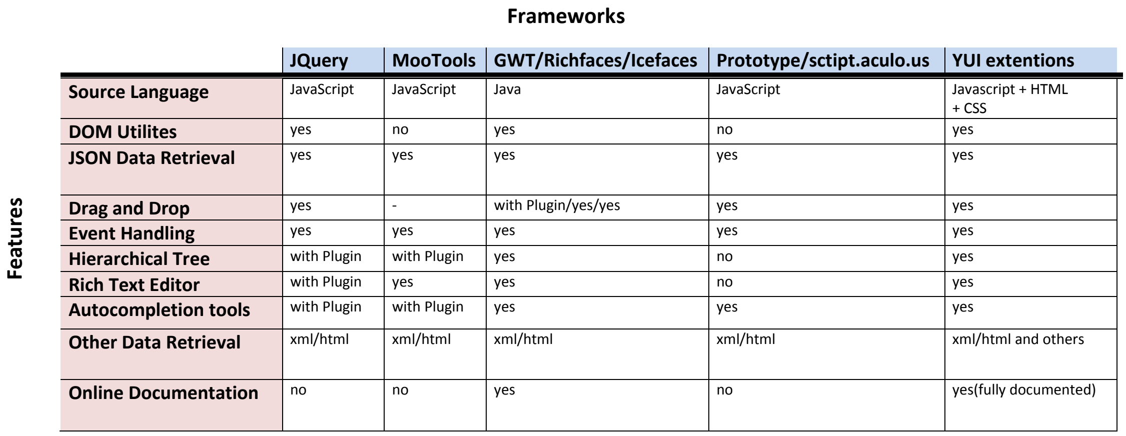
Most people getting started with JavaScript these days are faced with the challenging task of picking a JavaScript library to use. One of my dilemma was to choose a suitable JS library which works with PHP. There are frameworks out there that are worth digging into as well. Prototype, Script.aculo.us, Mootools, JQuery and YUI are all great choices. Which one you choose really has more to do with your own style and what you need to accomplish. The JS frameworks such Richfaces/IceFaces and GWT are excluded, since there are meant to work only with Java.

In order to complete two requirments from the objective 4 (see section 5.4) took me a lot of time to compare the features of each JavaScript library which are illustrated in figure 5.10. In this figure are picked only the features which should be used to complete the requirments. The libraries that fulfill all these requirment are JQuery and Yui. The two most interesting features are the Rich Text Editor and Hierarchical Tree Menu. In JQuery, Rich Text Editor can be only intergrated as plugin such as Lightweight Rich Text Editor [11]. Lightweight Rich Text Editor has only RTE/WYSIWYG editor's API, but no any controls/toolbars. The Rich Text Editor from Yui library has the standard controls/toolbars, including common structural treatments like lists, formatting treatments like bold and italic text, and drag-and-drop inclusion and sizing of images. The Rich Text Editor's toolbar is extensible via a plugin architecture so that advanced implementations can achieve a high degree of customization. The Hierarchical Tree Menu which is used to complete the second requirment, visualizing of consumed data from the web is an accordion menu. The accordation menu is a complex task and costs time to implement as plugin for JQuery. Therefore, I decided to use BubblingLibrary [12], an YUI extenstion which handles accordion menu. It handle the states of the items using JavaScript, you only need a reference to a DOM element within an accordion item, and you can modify the state for that particular item, collapsing or expanding it. Also you can control the accordion as a single object.

Development under jQuery takes more time to implememt these two requirments but it gives really good performance. YUI force us to follow predefined coding patterns and provide better maintenance. It is designed as framework to cover all the aspects of UI development as a holistic system. jQuery has a solid core of basic API, but is meant to extend by mixed plugins.

[11] RTE for JQuery http://archive.plugins.jquery.com/project/lwRTE

[12] BubblingLibrary:accordion menu, http://www.bubbling-library.com/themes/bubbling/jscripts/yui-cms/examples/accordion/manager.html



Frameworks

Features

	JQuery	MooTools	GWT/Richfaces/Icefaces	Prototype/sctipt.aculo.us	YUI extentions
Source Language	JavaScript	JavaScript	Java	JavaScript	Javascript + HTML + CSS
DOM Utilites	yes	no	yes	no	yes
JSON Data Retrieval	yes	yes	yes	yes	yes
Drag and Drop	yes	-	with Plugin/yes/yes	yes	yes
Event Handling	yes	yes	yes	yes	yes
Hierarchical Tree	with Plugin	with Plugin	yes	no	yes
Rich Text Editor	with Plugin	yes	yes	no	yes
Autocompletion tools	with Plugin	with Plugin	yes	yes	yes
Other Data Retrieval	xml/html	xml/html	xml/html	xml/html	xml/html and others
Online Documentation	no	no	yes	no	yes(fully documented)

Figure 5.10: Javascripts comparation based on their features

5.5.5 Technologies and Tools

5.5.5.1 ARC2

ARC2 is a flexible and a simple RDF system for semantic web applications and PHP practitioners. It is free, open-source, easy to use, and runs in most web server environments. It uses object-oriented code for its components and methods, but the processed data structures consist of simple associative arrays, which leads to faster operations and less memory consumption.

ARC is easy to combine with the existing PHP/MySQL software:

- works only with PHP4 and PHP5
- allows not use global variables, constants, or other practices which may cause integration problems,
- Components are kept as small as possible, only immediately required code is loaded into memory,
- Avoidance of PHP warnings and notices, error collection.
- Automatic path detection for component inclusion.
- Customizable per-store prefix for database tables.

The ARC2 has the following features and components:

- Various parsers (RDF/XML, Turtle, RSS 2.0, Google Social Graph API JSON)
- SemHTML RDF extractors (DC, eRDF, microformats, OpenID, RDFa)
- RDF Storage using MySQL (SPARQL SELECT, ASK, DESCRIBE, CONSTRUCT, + aggregates, LOAD, INSERT, and DELETE)
- SPARQL Endpoint Class (Setting of compliant SPARQL endpoints)
- Remote Store Class for querying remote SPARQL endpoints as if they were local stores
- Turtle templating and SPARQLScript

- Plugins and Triggers

In this section I will not go to explain in detail each component of ARC2. We will only concentrate on the components or methods which have been used for the practical implementation of SemTex project. For further information about the use of ARC2, visit the ARC2's Webpage http://arc.semsol.org.

SPARQL Endpoint Class: To create an SPARQL endpoint in ARC2 is simple. There is now a protocol-compliant endpoint class which can be used for HTTP-based data access. The listing 5.1 explains the use of SPARQL Endpoint Class in creating RDF stores, that can be access through SPARQL endpoint. MySQL will be used to store the RDF triples. Configuration options can be provided during any class instantiation. Dynamically loaded sub-components will inherit their caller's configuration. This allows you to specify certain settings once, without having to worry about them later.

```

/* ARC2 static class inclusion */
include_once(’path/to/arc/ARC2.php’);

/* MySQL and endpoint configuration */
$config = array(
  /* db */
  ’db_host’ => ’localhost’, /* optional, default is localhost */
  ’db_name’ => ’database’,
  ’db_user’ => ’user’,
  ’db_pwd’ => ’password’,

  /* store name */
  ’store_name’ => ’my_endpoint_store’,

  /* endpoint */
  ’endpoint_features’ => array(
    ’select’, ’construct’, ’ask’, ’describe’,
    ’load’, ’insert’, ’delete’,
    ’dump’ /* dump is a special command for streaming SPOG export */
  ),
  ’endpoint_timeout’ => 60, /* not implemented in ARC2 preview */
  ’endpoint_read_key’ => ’’, /* optional */
  ’endpoint_write_key’ => ’somekey’, /* optional */
  ’endpoint_max_limit’ => 250, /* optional */
);

/* instantiation */
$ep = ARC2::getStoreEndpoint($config);

if (!$ep->isSetUp()) {
  $ep->setUp(); /* create MySQL tables */
}

                  }
```

Listing 5.1: Setting up a SPARQL endpoint in ARC2

Remote Stores and Endpoints Class : ARC provides a RemoteStore component (starting with revision 2008-07-04 and based on Morten Frederiksen's excellent RemoteEndpointPlugin) which makes it possible to work with SPARQL (and SPARQL+) endpoints (almost) as if they were local store. This compontent will be used for accessing data from the Linked Data sources such as DBPedia or Gonames. The Listing 5.2 shows an example of remote store instantiation which in fact is the same as local store, but with a remote_store_endpoint configuration parameter (e.g. http://dbpedia.org/sparql) instead of database settings.

```

/* ARC2 static class inclusion */
include_once(’path/to/arc/ARC2.php’);

/* configuration */
$config = array(
  /* remote endpoint */
  ’remote_store_endpoint’ => ’http://dbpedia.org/sparql’,
);

/* instantiation */
$store = ARC2::getRemoteStore($config);
```

Listing 5.2: Setting up a remote SPARQL endpoint in ARC2

To run a query on remote SPARQL endpoint the query method will be used as it listed in 5.3:

```

$query = ’PREFIX p: <http://dbpedia.org/property/>
                  SELECT DISTINCT ?label WHERE {
                     <http://dbpedia.org/resource/’.$searchTerm.’> p:redirect ?label.
                                   }’;

$rows_name = $store->query($query, ’rows’);
```

Listing 5.3: Running an SPARQL query

5.5.5.2 Zend Gdata

Zend_Gdata is a component of Zend Framework [13]. Zend Gdata is about Google Data. Google Data APIs provide programmatic interface to some of Google's online services. The Protocol of GData is based upon the Atom Publishing Protocol [14] and allows client applications to retrieve data matching queries, post data, update data and delete data using standard HTTP and the Atom syndication formation. The Google data Protocol is based upon the Atom Publishing Protocol and allows client applications to retrieve data matching queries, post data, update data and delete data using standard HTTP and the Atom syndication formation. The Zend_Gdata component was written entirely in PHP 5. It will not run on any server that does not have a minimum of PHP 5.1.4 installed. The Zend_Gdata component also supports accessing other services implementing the Atom Publishing Protocol.

The online services that are available include the following

- **Google Calendar** is a calendar application
- **Google Spreadsheets** provides a collaborative spreadsheets tool to store the data for your applications.
- **Google Documents List** provides a list of all spreadsheets, word processing documents, and presentations stored in a Google account.
- **Google Provisioning** provides the functionality to manipulate user accounts, nickname and email lists on a Google Apps hosted domain.
- **YouTube** provides the functionality to search and retrieve videos, comments, favorites, subscriptions, user profiles and more.
- **Google Base** provides the functionality to retrieve, post, update, and delete things in Google Base.
- **Picasa Web Albums** provides an online photo sharing application.
- **Google Blogger** is a popular Internet provider of push-button publishing and syndication.
- **Google Notebook** allows to show public notebook content
- **Google CodeSearch** allows you to search public source code from many projects.

YouTube is one of services used in SemTex Project. The YouTube Data API provides ability to read and write to YouTube's content. Users can perform unauthenticated requests to Google Data feeds to retrieve feeds of popular videos, comments, public information about YouTube user profiles, user playlists, favorites, subscriptions and so on. The YouTube Data API allows read-only access to public data, which does not require authentication. For any write requests, a user needs to authenticate either using ClientLogin or AuthSub authentication [15].

5.5.5.3 SOAP

SOAP (Simple Object Access Protocol) is a lightweight XML-based protocol for exchanging structured and typed information between peers in a decentralized, distributed environment over native web protocols, such as HTTP. The first version 1.1 of protocol has been released in 2000 and became a W3C recommendation. The current version is 1.2 and was released in 2007.

[13] Zend Framework, http://framework.zend.com

[14] Atom Publishing Protocol, http://ietfreport.isoc.org/idref/draft-ietf-atompub-protocol/

[15] AuthSub, http://code.google.com/apis/youtube/developers_guide_php.html#Authentication

SOAP is fundamentally a stateless, one-way message exchange paradigm, but applications can create more complex interaction patterns (e.g., request/response, request/multiple responses, etc.) by combining such one-way exchanges with features provided by an underlying protocol and/or application-specific information. SOAP is silent on the semantics of any application-specific data it conveys, as it is on issues such as the routing of SOAP messages, reliable data transfer, firewall traversal, etc. However, SOAP provides the framework by which application-specific information may be conveyed in an extensible manner. Also, SOAP provides a full description of the required actions taken by a SOAP node on receiving a SOAP message.

SOAP uses structured envelope as it listed in 5.4 for sending Web Services messages over HTTP.

```

<soapenv:Envelope xmlns:soapenv=\”http://schemas.xmlsoap.org/soap/envelope/\”
    xmlns:urn=\”urn:TextAnalyzer\”>
  <soapenv:Header>
  <soapenv:Header/>
  <soapenv:Body>
    <urn:analyze>
     <urn:text>$text</urn:text>
    </urn:analyze>
  </soapenv:Body>
</soapenv:Envelope>
```

Listing 5.4: An example of a SOAP message from SemTex project.

The SOAP envelope consists of two parts:

1. An optional header providing information on authentication, encoding of data, or how a recipient of a SOAP message should process the message.

2. The body that contains the message. These messages can be defined using the WSDL specification.

The Web Services Description Language (WSDL) provides a model and an XML format for describing Web services that use SOAP protocol. The Listing 5.5 shows the WSDL definition of a text analysis service. This service supports a single operation called *analyze*, which is deployed using the SOAP protocol over HTTP. *analyze* operation request takes a text of type string, and returns the the extracted terms as a string.

```
<?xml version="1.0" encoding="UTF-8"?>
<definitions name="TextAnalyzer"
 targetNamespace="http://inmw016.joanneum.at/TextAnalyzerSoapService"
 xmlns:tns="http://inmw016.joanneum.at/TextAnalyzerSoapService"
 xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
 xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema"
 xmlns:ns="urn:TextAnalyzer"
 xmlns:SOAP="http://schemas.xmlsoap.org/wsdl/soap/"
 xmlns:MIME="http://schemas.xmlsoap.org/wsdl/mime/"
 xmlns:DIME="http://schemas.xmlsoap.org/ws/2002/04/dime/wsdl/"
 xmlns:WSDL="http://schemas.xmlsoap.org/wsdl/"
 xmlns="http://schemas.xmlsoap.org/wsdl/">

<types>
 <schema targetNamespace="urn:TextAnalyzer"
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:TextAnalyzer"
  xmlns="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified"
  attributeFormDefault="unqualified">
  <import namespace="http://schemas.xmlsoap.org/soap/encoding/" schemaLocation="http://schemas.xmlsoap.org/soap/encoding/"/>
  <!-- operation request element -->
  <element name="analyze">
   <complexType>
    <sequence>
     <element name="text" type="xsd:string" minOccurs="1" maxOccurs="1"/>
    </sequence>
   </complexType>
  </element>
  <!-- operation response element -->
  <element name="analyzeResponse">
   <complexType>
    <sequence>
     <element name="result" type="xsd:string" minOccurs="1" maxOccurs="1"/>
    </sequence>
   </complexType>
  </element>
 </schema>
</types>

<message name="analyzeRequest">
 <part name="parameters" element="ns:analyze"/>
</message>

<message name="analyzeResponse">
 <part name="parameters" element="ns:analyzeResponse"/>
</message>

<portType name="TextAnalyzerPortType">
 <operation name="analyze">
  <documentation>text</documentation>
  <input message="tns:analyzeRequest"/>
  <output message="tns:analyzeResponse"/>
 </operation>
</portType>

<binding name="TextAnalyzer" type="tns:TextAnalyzerPortType">
 <SOAP:binding style="document" transport="http://schemas.xmlsoap.org/soap/http"/>
 <operation name="analyze">
  <SOAP:operation soapAction=""/>
  <input>
     <SOAP:body parts="parameters" use="literal"/>
  </input>
  <output>
     <SOAP:body parts="parameters" use="literal"/>
  </output>
 </operation>
</binding>

<service name="TextAnalyzer">
 <documentation>gSOAP 2.7.13 generated service definition</documentation>
 <port name="TextAnalyzer" binding="tns:TextAnalyzer">
  <SOAP:address location="http://localhost/cgi-bin/TextAnalyse/at_server.exe"/>
 </port>
</service>
</definitions>
```

Listing 5.5: An example of WSDL definition that use SOAP protocol.

5.5.5.4 REST

Besides SOAP there is an alternative for the realization of Web services called REST architecture. Representational State Transfer (REST) is a software architectural style for distributed hypermedia systems like the world wide web. REST is a term introduced by Roy Fielding in his Ph.D. dissertation ”Architectural Styles and the Design of Network-based Software Architectures”.

Roy Fielding explained the meaning of Representational State Transfer as follows [Fielding, 2000]:

”Representational State Transfer is intended to evoke an image of how a well-designed Web application behaves: a network of web pages (a virtual state-machine), where the user progresses through an application by selecting links (state transitions), resulting in the next page (representing the next state of the application) being transferred to the user and rendered for their use.”

REST is a managed set of architectural constraints that attempts to minimize latency and network communication, while at the same time maximizing the independence and scalability of component implementations. REST enables the caching and reuse of interactions, dynamic substitutability of components, and processing of actions by intermediaries, in order to meet the needs of an Internet-scale distributed hypermedia system [Fielding and Taylor, 2002].

The REST interface is generic. There are no protocol conventions that must be known so that client and server can communicate. The following list describes the meaning of the HTTP methods used by REST:

1. **GET**: GET queries the representation of a resource. GET requests can be safely sent. One can not draw a client for its impact in the responsibility.

2. **POST**: POST add new/update item in a specific resource. POST is not free from side effects. For example, a POST call changes fields in a database or starts processes are on the server.

3. **PUT**: New resources can be created or the content of existing resources can be replaced using PUT method.

4. **DELETE**: Resources can be deleted using DELETE method.

Each REST Resource has a generic interface over the HTTP methods GET, POST, PUT and DELETE. With these four methods, most applications can be covered. Many applications that use SQL apply only the generic commands SELECT, INSERT, UPDATE, and DELETE.

One of the most interesting aspect of the REST is the security angle. Since the REST massages go over HTTP or HTTPS, the administrator (or firewall) can discern the intent of each message by analyzing the HTTP command used in the request. For example, a GET request can always be considered safe because it can not, by definition, modify any data. It can only query data.

According to the REST security, many web services of companies are implemented using the REST architecture. Examples are Google, Flickr, Facebook which have implemented most of thier web services using the REST architecture.

5.5.5.5 Virtuoso TS

See section 3.4.1.1.

5.5.5.6 TDB Jena over Joseki

See section 3.4.1.2

5.5.5.7 YUI framework

The YUI[16] (Yahoo! User Interface) framework is a set of utilities and controls, written with JavaScript and CSS, for building richly interactive web applications using techniques such as DOM scripting, DHTML AJAX and JSON. All components in the YUI framework have been released as open source under a BSD license and are free for all users. It has been well-developed by Yahoo so that it can be adapted in different ways. One of the adaptations of YUI is to create an Ajax based application. With YUI one can create not only web based applications but also widgets that can easily be implemented in a desktop.

The YUI framework consist of three main components:

1. **Utilities**: they simplify in-browser development that relies on cross-browser DOM scripting, as do all web applications with DHTML and AJAX characteristics.

2. **UI Controls**: provide highly interactive visual design elements for your web pages. These elements are created and managed entirely on the client side and never require a page refresh.

3. **CSS resources**

YUI advantages:

Great Documentation: Yahoo made a great step towards ensuring that the library is understood and well documented. From method signatures to parameters and API everything is well thought of by Yahoo. *Good Support* : Only the name Yahoo can convince developers about the program. Also, there are many people who use this library, and have enough experience to help.

DOM Event Handling: DOM can easily be implemented as it uses element IDs. Usually DOM could be seen passing through the node but not in YUI. This is a good thing since it prevents different activities from mixing up. It can be considered as a straightforward command. No need for different nodes which could be confusing for different browsers

YUI disadvantages:

Issues with Opera: apparently YUI has trouble implementing its programs in this browser. If you are thinking of developing a widget or an Ajax based program for Opera, better think of other libraries or at least use it as a program. Different developers have reported YUI's inefficiency in Opera especially when running a widget.

Additional HTML div tag required: this situation is applicable especially in widgets. I happen to categorize this as bad since programs can easily be created in different libraries. Creating widgets using YUI needs additional customization. Although it gives you a freehand in developing widgets up to the last minute, it is going to be a challenge for starting developers. Better have a good knowledge in JavaScript if you are interested in building widgets from YUI.

HTML Modification: In relation to HTML div tag requirements, HTML should be further configured so that it could work well in different browsers (except Opera). That is another challenge considering HTML should still be configured to work with HTML div tag.

In conclusion, YUI is proven to be scalable, fast, and robust. It has lovely syntax, fantastic documentation, is increasingly dominant and well-supported, is lightning fast, feels very ”clean”, largely keeps itself to itself, and the developers definitely know what they're doing. Built by frontend engineers at Yahoo! and contributors from around the world, it is an industrial-strength JavaScript library for professionals who love JavaScript.

[16] yui, http://developer.yahoo.com/yui/

5.6 Linked Datasets

A basic entity in Linked data sets is dataset, that can be defined as:

"A dataset is a set of RDF triples that are published, maintained or aggregated by a single provider." [Alexander et al., 2009].

Dataset as a collection of RDF triples deals with structured information from specific source (e.g. Wikipedia) that are hosted on a specific server (e.g. DBPedia). It can be accessed on the Web, for example through resolvable HTTP URIs or through a SPARQL endpoint.

The idea of linked datasets was firstly introduced by Sir Tim Berners-Lee, provided by a set of rules [17]. It is all about making links , so that humans or machines can easily find other, related, data on the Web of Data. And so has arisen the idea of creating the Linking Open Data (LOD [18]) community project as a collaborative effort applying the linked data rules. It aims at bootstrapping the Semantic Web by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [Bizer et al., 2007a]. The number of datasets in LOD commnunity is always growing, but still there is a lack of information in datasets in building Linked data applications. DBpedia and Geonames are the only two highly interlinked datasets and contain sufficient information in achieving the practical part of this thesis.

5.6.1 DBPedia

DBpedia [19] is a community effort of an initiative aiming at extracting structured information from Wikipedia and to make this information available on the Web. It allows to use Semantic Web techniques that can be employed against it asking sophisticated queries against Wikipedia, linking it to other datasets on the Web according to Tim Berners-Lee's Linked Data principles, or creating new applications [Auer et al., 2008].

The DBpedia project made the following contributions to the development of the Web of Data[Auer et al., 2008]:

- An information extraction framework has been developed, which converts Wikipedia content to RDF.
- DBpedia provides Wikipedia content as multi-domain datasets which consists of more than 103 million RDF triples.
- DBpedia dataset is interlinked with other open datasets. This results in a large Web of data containing altogether around 2 billion RDF triples.
- A series of interfaces and access modules has been developed which allow other third party applications to access the dataset via Web services and linked to other sources.

The DBpedia dataset currently provides information about more than 3.4 million entities out of which 1.5 million are classified in a consistent Ontology, including at least 312.000 persons, 413.000 places, 140.000 organizations, 94.000 music albums, 49.000 films, 146.000 species and 4.600 diseases. It also contains 841.000 links to images, 5.081.000 links to external web pages, 9.393.000 external links into other RDF datasets, 565.000 Wikipedia categories and 75.000 YAGO [20] categories. The DBpedia altogether consists of over 1 billion RDF triples out of which 257 million were extracted from the English version of Wikipedia and 766 million were extracted from other language versions including German,French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish

[17] Linked Data Sets, http://www.w3.org/DesignIssues/LinkedData.html

[18] LOD, http://linkeddata.org/

[19] DBPedia, http://wiki.dbpedia.org/About

[20] Yago, http://www.mpi-inf.mpg.de/yago-naga/yago/

and Norwegian. Every resource in DBpedia dataset is described by a label, a short abstract, a long abstract, links to other sources and links to images (if available). These are the basic information which can be retrieved from DBPedia dataset. The table 5.1 shows the numbers of abstracts and labels per language.

Table 5.1: Numbers of labels and abstracts per language (March 2010)

Language	Number of Labels	Number of Abstracts
English	7,311.000	3,144,000
German	589,500	503,000
French	653,200	545,000
Polish	469,500	430,000
Dutch	433,600	392,000
Italian	487,700	381,000
Spanish	445,000	362,000
Japanese	288,200	275,000
Portuguese	428,300	367,000
Swedish	266,400	213,000
Chinese	182,100	179,000

The DBpedia dataset can be either imported as RDF dumps into third party applications or can be accessed online using variuos user interfaces of DBpedia. These DBpedia user interfaces currently use OpenLink Virtuoso [21] and MySQL as storage back-ends.

5.6.2 Geonames

GeoNames[22] is a geographical database that is licensed under a Creative Commons attribution licence [23]. Database contains over 8 million names of places in different languages and 7 million unique features out of which 2.6 million populated places and 2.8 million alternate names. All these feature are catagorized into 9 feature classes and further subcategorized into 645 feature codes. The alternate names can be usually used in name disambiguation (e.g Vienna in United Stated and Vienna in Austria). The data set include geographical latitude and longitude data that can used to find images or videos made in specific location. The developers have a free access to the data through a number of webservices. They can also download the database and set up their own database with the data and query on it locally.

Geonames provides also RDF descriptions of millions of geographical locations worldwide. Over 6.2 million geonames toponyms now have a unique URL with a corresponding RDF web service. Geonames offers the following Entry Points to access the GeoNames Semantic Web:

- navigation via mother earth [24] to discover Linked Data links
- use the geonames search webservice with the type=rdf parameter option
- download the database dump and construct the url for the features using the pattern ”http://sws.geonames.org/ge
- RDF dump with 6.520.110 features and 93.896.732 triples. The dump has one rdf document per toponym on every line of the file.

[21] OpenLink Virtuoso, http://virtuoso.openlinksw.com/

[22] Geonames, http://www.geonames.org/

[23] Creative Commons, http://creativecommons.org/

[24] Mother earth, http://www.geonames.org/6295630/earth.html

If we compare Geonames with DBPedia, we note that Geonames does not provide a SPARQL Endpoint. Even if we download the RDF dump, it is hard to extract RDF document using SPARQL engines. Therefore I decided to download the entire SQL database in order to achieve the goal of this work.

5.7 Implementation

In the previous sections have been introduced the concept, architecture and the use cases of SemTex application. This section presents more in depth detail each component of SemTex prototype. The whole process of this application uses techniques like extraction, classification, interlinking with distinct open data sources from Linked Data and publishing. Corresponding to figure 5.11 depicting the implemented work flow will explained the phases by the sequence of their appearance. In the first phase a text must be edited by an online editor. In the second phase the posted text is analyzed for interesting terms. The classified terms are sent to third phase which automatically suggests further information from the Web of Data. This includes data from various Linked Data sources such as textual content, links, images, or video. In the fourth phase the edited text interlinked with further information from the Web will be published.

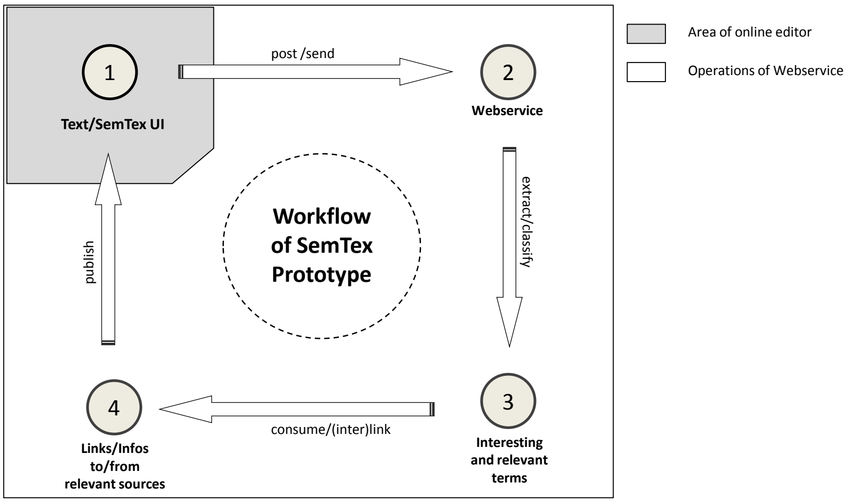


Figure 5.11: Workflow of SemTex prototype

5.7.1 Configuration

Before starting the SemTex application, we should take into account the tools that must be installed and configured. The SemTex application require the following technologies:

- Web Server - Apache 2.2.11
- Database Server- MySql 5.1.36
- Server side scripting language - PHP 5.3.0
- RDF database server - Virtouso 6.1.0
- RDF database server - TDB Jena 0.8.2
- RDF Web Server - Joseki 3.4.0

The SemTex application has been developed using the operating system Windows XP since all the mentioned technologies can be easily installed. The first three technologies are installed and configured using a powerful project named WAMP [25]. WAMP is an acronym formed from the initials of the operating system Microsoft Windows and the principal components of the package: Apache, MySQL and PHP. After WAMP has been installed the directory of SemTex project including php files is moved to a ”www” directory (generally c:\wamp\www). But before testing the project on the localhost (generally http:\\localhost\SemTex) the SOAP module as a PHP extension must be manually configured in order to enable the communication with the term extraction module. Then the executable file including the WSDL file of term extraction module are moved to the directory path wamp\bin\apache\Apache2.2.11\cgi-bin. SemTex require also a geoname database which is configured using the MySql database server by setting an adequate username, password, database name and host.

Finally, the project can be started on localhost. The DBpedia dataset can be accessed via HTTP URIs and queries on remote SPARQL endpoint [26] that is made available (through the Virtuoso instance hosting Linked Open Data). Response times for queries depend on server load and it could happen that a resource is unavailable for any (technical) reason that can not be influenced by the data consumer. One of solution to this issue is hosting a copy of DBPedia dataset on infrastructure that is under one’s own control. Therefore, the Virtouso and TDB Jena server has been installed and configured and contain the DBPedia dataset which can be accessed using a local SPARQL Endpoint. Joseki is a web server that supports the HTTP binding for the SPARQL protocol and a protocol for SPARQL/Udpate in order to influence the RDF data from TDB Jena server. These both RDF databases perform the same task. The reason of installing of these RDF databases was to test the response time for queries and also the load time of datasets.

5.7.2 Term Extraction and Classification

In Semantic Web applications, the text analysis including term recognition and extraction are tasks of paramount importance. The manual text analysis of these tasks is laborious and therefore cost-intensive, and would profit from a maximum level of automation. For this purpose, the identification and extraction of terms that play an important role in the domain under consideration, is a vital first step [Maynard et al., 2008].

Based on the existing techniques, methods and tools of terminolgy extraction, as mentioned in chapter 2, SemTex Web application has included term extraction module, which is targeted at German-speaking content as the cooperating content provider produces the majority of its content in German.

[25] WAMP, http://www.wampserver.com

[26] DBPedia SPARQL endpoint, http://dbpedia.org/sparql

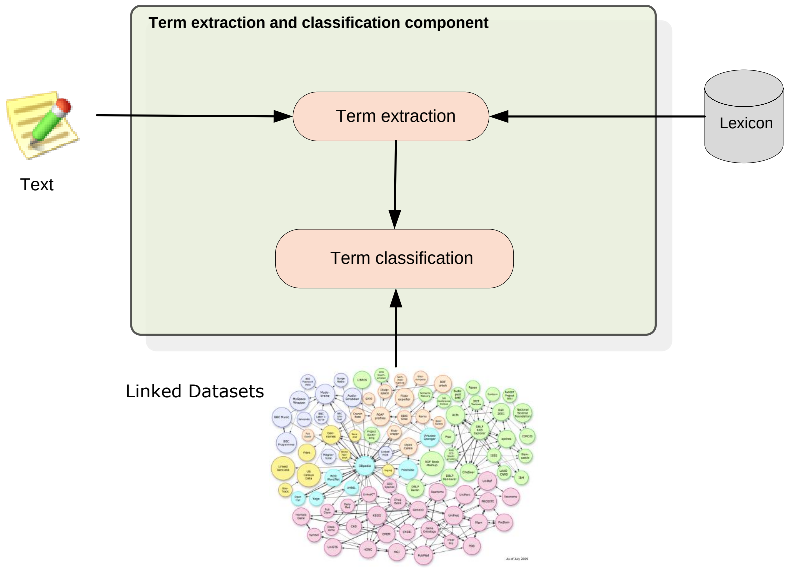


Figure 5.12: Term extraction and classification component

Even though there exist several term extraction solutions for the English language (which can also be plugged in and used in the SemTex Web application) there is still a lack of well-performing, generally available tools for German. Through its modular structure the Semtex Web aplication can support term extraction solutions for any language that are available from third-party providers. The term extraction module work closely with the classification module which was one of the goal of this thesis. The figure 5.12 shows the modules of Term extraction and classification component and their interactions. These module are explained in the following subsections.

5.7.2.1 Term Extraction

The developed method of term extraction in the SemTex project is based on a simple recipe: all terms and possible word formations that occur in a general dictionary are removed from the text. As a result, this step delivers ”interesting” words such as names, toponyms, and domain-specific terms.

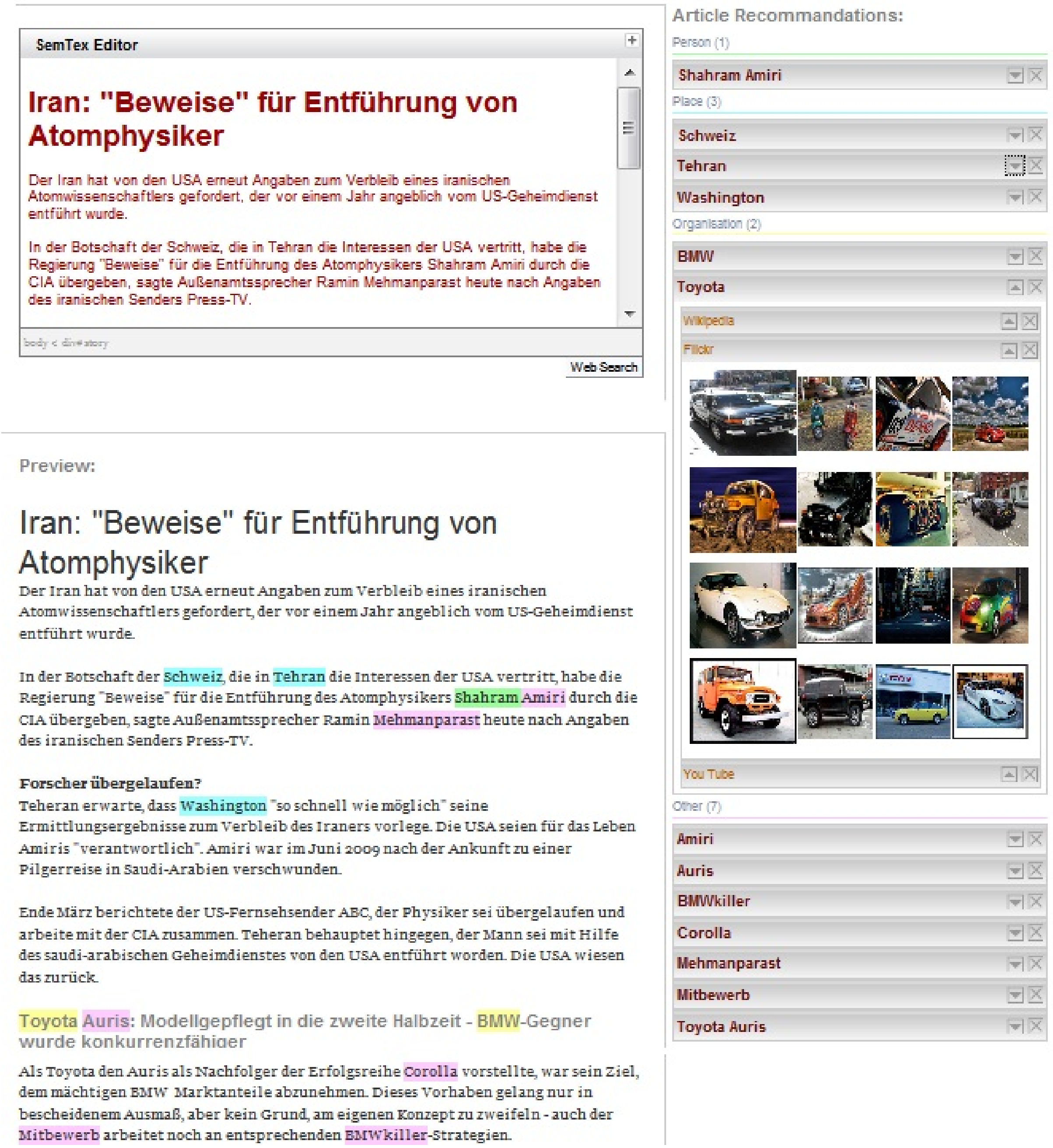
In order to achieve this approach all possible morphological formations have to be considered. Therefore, it is required a complete dictionary with all the morphological forms. Since the Institute for Information at JOANNEUM Research [27] works already together with the Bertelsmann publishing group in other projects; it was possible to use for this research project, a full word list of ”Wahrig” [28] german dictionary. From this word list was generated a full form lexicon that is stored in a compact form (as a minimal finite automaton).

This can perform the following queries efficiently:

- Exact query of a word

[27] JOANNEUM Research, http://www.joanneum.at

[28] Wahrig Dictionary, http://www.babylon.com/dictionary/wahrig/german/



Iran: "Beweise" für Entführung von Atomphysiker

Der Iran hat von den USA erneut Angaben zum Verbleib eines iranischen Atomwissenschaftlers gefordert, der vor einem Jahr angeblich vom US-Geheimdienst entführt wurde.

In der Botschaft der Schweiz, die in Tehran die Interessen der USA vertritt, habe die Regierung "Beweise" für die Entführung des Atomphysikers Shahram Amiri durch die CIA übergeben, sagte Außenamtssprecher Ramin Mehmanparast heute nach Angaben des iranischen Senders Press-TV.

Iran: "Beweise" für Entführung von Atomphysiker

Der Iran hat von den USA erneut Angaben zum Verbleib eines iranischen Atomwissenschaftlers gefordert, der vor einem Jahr angeblich vom US-Geheimdienst entführt wurde.

In der Botschaft der Schweiz, die in Tehran die Interessen der USA vertritt, habe die Regierung "Beweise" für die Entführung des Atomphysikers Shahram Amiri durch die CIA übergeben, sagte Außenamtssprecher Ramin Mehmanparast heute nach Angaben des iranischen Senders Press-TV.

Forscher übergelaufen?
Teheran erwarte, dass Washington "so schnell wie möglich" seine Ermittlungsergebnisse zum Verbleib des Iraners vorlege. Die USA seien für das Leben Amiris "verantwortlich". Amiri war im Juni 2009 nach der Ankunft zu einer Pilgerreise in Saudi-Arabien verschwunden.

Ende März berichtete der US-Fernsehsender ABC, der Physiker sei übergelaufen und arbeite mit der CIA zusammen. Teheran behauptet hingegen, der Mann sei mit Hilfe des saudi-arabischen Geheimdienstes von den USA entführt worden. Die USA wiesen das zurück.

Toyota Auris: Modellgepflegt in die zweite Halbzeit - BMW-Gegner wurde konkurrenzfähiger

Als Toyota den Auris als Nachfolger der Erfolgsreihe Corolla vorstellte, war sein Ziel, dem mächtigen BMW Marktanteile abzunehmen. Dieses Vorhaben gelang nur in bescheidenem Ausmaß, aber kein Grund, am eigenen Konzept zu zweifeln - auch der Mitbewerb arbeitet noch an entsprechenden BMWkiller-Strategien.

Figure 5.13: Preview of term extraction

- All words with a specific prefix
- Search word, which contains placeholder for a character or contains any number of characters

Unfortunately, the German language has a property that has a negative impact on this algorithm: throught the word composition (composite words) can be generated almost any kind of long terms (multi-word terms), which usually are not in the dictionary. For this reason, a decompounding was integrated into the algorithm.

Two variants were developed:

1. A fast but imprecise method that always try to find the longest possible part of a word in the word list without using additional grammatical knowledge.

2. A slower but more accurate method that uses only the grammatical forms of words allowed in the decomposition

This method of term extraction does not work for proper names, which are homonymous to general terms (”Claudia Schmied”) or are combinations of such words(Salz-Burg, Klagen-Furt). This can be improved either by integration of multiple word lists (geographical concepts, first names, company names) as ”positive list”, or by a rule-based approach. A detailed implementation of this module is out of scope of this thesis. The figure 5.13 depicts a news article with color-highlighted terms extracted using the term extraction module. The color codes are used for term classification. The next section explains in details, how these terms can be filtered and classified using rule-based approach.

5.7.2.2 Term Classification

In order to classify the terms extracted from a german text and also filter some terms extracted with no meaning (such as ”BMWkiller” on the figure 5.13), the DBpedia dataset from Linked Data is used. The DBpedia dataset currently provides information about more than 3.4 million ”things”, about 1.5 million are classified in a consistent Ontology, including 312,000 persons, 413,000 places (including 310,000 populated places), 140,000 organizations and others.

The classification of these terms is done via a SPARQL queries to DBPedia as illustrated in listing 5.6, 5.7, 5.8.

```

PREFIX rdf:         <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

        ASK WHERE {
                     <http://dbpedia.org/resource/Graz> rdf:type ?typePlace .
                     FILTER (sameTerm(?typePlace,dbpedia-owl:Place))
                  }
```

Listing 5.6: A SPARQL query for classification of places

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

        ASK WHERE {
                     <http://dbpedia.org/resource/IBM> rdf:type ?typeOrg .
                     FILTER (sameTerm(?typeOrg,dbpedia-owl:Organisation))
                  }
```

Listing 5.7: A SPARQL query for classification of organisations

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

        ASK WHERE {
                     <http://dbpedia.org/resource/Barack_Obama> rdf:type ?
                         typePerson .
                     FILTER (sameTerm(?typePerson,dbpedia-owl:Person))
                  }
```

Listing 5.8: A SPARQL query for classification of persons

Each of the 1.5 million classified resources in the DBpedia dataset is identified by a URI reference of the form http://dbpedia.org/resource/Name, where term Name is taken from the SemTex Web application. rdf:type is an instance of rdf:Property that is used to state that a resource is an instance of a class. In listings above an ASK query is used to check whether the resource (e.g. http://dbpedia.org/resource/Barack_Obama) is a type of an specific class (e.g. Person) from the dbpedia ontology (dbpedia-owl). ASK query returns a boolean indicating whether a query pattern matches or not.

Before asking for term classification, we must take under consideration that the DBPedia uses URI reference that works only in english. This thesis is mostly focused at finding information in german language. Therefore, asking for the german terms (e.g. Wien) in DBPedia require another SPARQL query as shown in listing 5.9. The redirect property points to another thing (e.g. Vienna) which contains informations in german language.

```

PREFIX p: <http://dbpedia.org/property/>

        SELECT DISTINCT ?resource WHERE {
               <http://dbpedia.org/resource/Wien> p:redirect ?resource .
        }
```

Listing 5.9: A SPARQL query with a redirect property

5.7.3 Linked Data consumtion and Interlinking

The description of Linked Data consumtion and Interlinking component as central scope of the practical implementations of this thesis including its sub-modules and tools is followed in this section. The figure 5.14 depicts the implemented modules within Linked Data consumtion and Interlinking component. As we can see from the figure 5.14 the component contains only two modules from Linked Data, DBpedia and GeoNames, two datasets that already contain a lot of general knowledge and geographic information. Even though the use of further datasets is easily possible, this first step aims at identifying relevant concepts for the found terms where these two Linked Data hubs already provide sufficient information for general news articles.

Once the concept has been identified it is possible to retrieve further information from different data sources. The information gained from Linked Data sources can also be extremely useful for queries to other repositories containing for instance user contributed multimedia content. It is possible to find more appropriate matches as the query can be enhanced with more metadata for finding relevant images and videos. Especially for concepts related to geographical locations are implemented two other Web API 2.0 modules such as Flickr and YouTube. Web 2.0 APIs are a well-known alternative to linked datasets. It is possible to supply coordinates that have been retrieved from DBpedia or GeoNames in the query when searching for images on Flickr or videos on Youtube.

In many cases it is not possible to find only one distinct concept but rather many potential matches are found and thus disambiguation needs to be done. In the current use case there is almost no information about the term available directly as it is simply part of a news article. The information about the category of the article (e.g. politics, local news, etc.) can aid in narrowing potential concepts though. Subsequently this implies that link discovery and resolution approaches that have been proposed for Linked Data sources such as SILK (see section 3.5.2), the OOD-Linker (see section 3.5.3) or KnoFuss (see section 3.5.4) cannot be used in this context. These approaches rely on further information about a resource that can be compared in two different datasets, i.e. there needs to be some overlap between the source and target datasets which is not the case in our scenario where the source dataset contains no more information about a concept than its label.

5.7.3.1 DBpedia module

DBpedia module is implemented based on the structured information from Wikipedia which are available on the Web. These information were converted into structured knowledge so that by using the Semantic Web technologies can be created various queries against Wikipedia in order to build Web application such as SemTex.

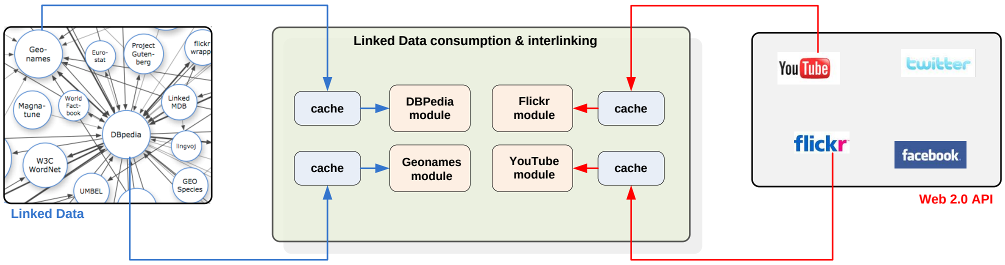
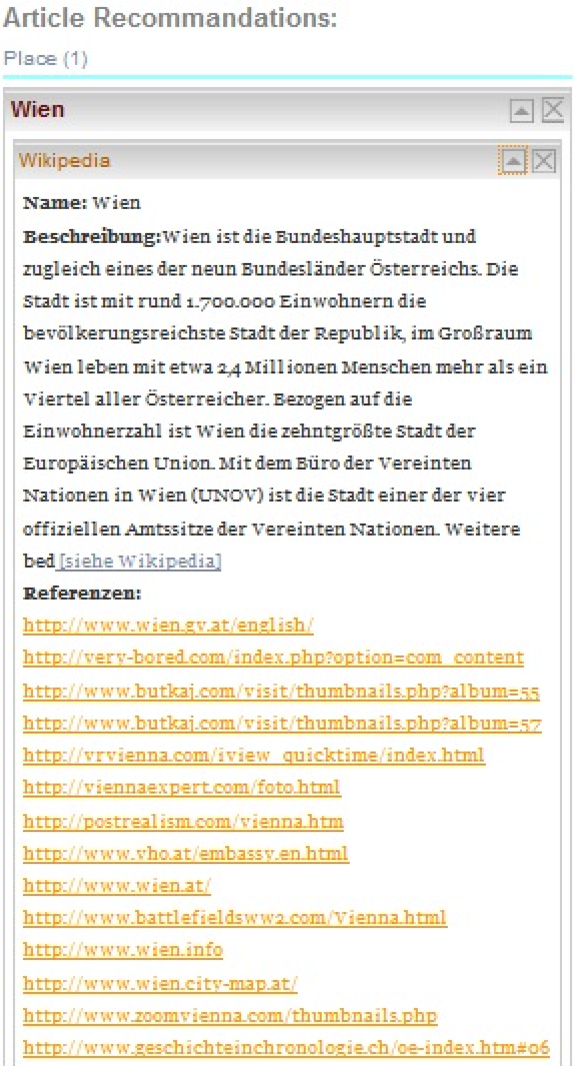


Figure 5.14: Linked Data consumption and Interlinking Component

This module receive as input the terms extracted from the text (typed by an editor) which are targeted against DBpedia. The DBpedia results with specific information based on the terms queried which is passed to the SemTex web application presenting the recommendations to the editor (see figure 5.15).



Article Recommandations:

Place (1)

Wien

Wikipedia

Name: Wien

Beschreibung: Wien ist die Bundeshauptstadt und zugleich eines der neun Bundesländer Österreichs. Die Stadt ist mit rund 1.700.000 Einwohnern die bevölkerungsreichste Stadt der Republik, im Großraum Wien leben mit etwa 2,4 Millionen Menschen mehr als ein Viertel aller Österreicher. Bezogen auf die Einwohnerzahl ist Wien die zehntgrößte Stadt der Europäischen Union. Mit dem Büro der Vereinten Nationen in Wien (UNOV) ist die Stadt einer der vier offiziellen Amtssitze der Vereinten Nationen. Weitere bed [siehe Wikipedia]

Referenzen:

http://www.wien.gv.at/english/
http://very-bored.com/index.php?option=com_content
http://www.butkaj.com/visit/thumbnails.php?album=55
http://www.butkaj.com/visit/thumbnails.php?album=57
http://vrvienna.com/iview_quicktime/index.html
http://viennaexpert.com/foto.html
http://postrealism.com/vienna.htm
http://www.vho.at/embassy.en.html
http://www.wien.at/
http://www.battlefieldsww2.com/Vienna.html
http://www.wien.info
http://www.wien.city-map.at/
http://www.zoomvienna.com/thumbnails.php
http://www.geschichteinchronologie.ch/oe-index.htm#06

Figure 5.15: Term-Related Information from Wikipedia

The main problem of the term search in DBpedia, in general, is that the results are obtained merely on the basis of a syntactic match (i.e. the exact occurrence of a given term). So that asking for some information from DBpedia may result to the wrong or empty information. So that using the existing knowledge provided by DBpedia the following spelling problems are solved:

- Synonyms
- Alternative spelling (e.g. Barack Obama vs. Obama)
- Translations in different languages (e.g. Vienna vs. Wien)

All these problems are solved using remote DBpedia SPARQL endpoint. But before sending queries to remote DBpedia SPARQL endpoint the ARC library must be included, which is discribed in section 5.5.5.1. The ARC library provides a RemoteStore component which makes it possible to work with SPARQL endpoints as if they were local stores. The SPARQL endpoint configuration and remote Store instantiation of DBpedia is shown in listing 5.10.

```
    /* configuration */
    $config = array(
    /* remote endpoint */
    //
        'remote_store_endpoint' => "http://dbpedia.org/sparql",
    );

    /* instantiation */
    $store = ARC2::getRemoteStore($config);
```

Listing 5.10: The SPARQL endpoint configuration and remote Store instantiation

The listing 5.11 illustrates an example of SPARQL query to solve a translation problem of a term. The property dbprop:redirect of the domain ontology extracted from DBpedia is the key of solving the problems with synonyms, alternative spelling and translations in different languages. Each thing in the DBpedia data set is identified by a URI reference of the form http://dbpedia.org/resource/Name, where Name represents the term extracted. In this case, the http://dbpedia.org/resource/Wien is redirected (using property dbprop:redirect) to http://dbpedia.org/resource/Vienna (which is the result of this query). The same query is also applied to solve the rest of problems (Synonyms and alternative spelling).

```

$query = 'PREFIX p: <http://dbpedia.org/property/>

                    SELECT DISTINCT ?label WHERE {
                       <http://dbpedia.org/resource/Wien> p:redirect ?label.
                    }';
// result
$row = $store->query($query, 'rows');
$result = $row[0]['label']; //  $result = "http://dbpedia.org/resource/Vienna"
```

Listing 5.11: A SPARQL query for term identification

Once the terms has been identified it is possible to retrieve further information from DBpedia source. The SPARQL query in listings 5.12,5.13 results to the information that fulfill the user's interests. These information include the long description and external links of the search terms. The first query uses also langMatches and lang to find the german abstracts for the term "Vienna".

```

PREFIX p: <http://dbpedia.org/ontology/>
                PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
                         SELECT DISTINCT ?abstract WHERE {
                              <http://dbpedia.org/resource/Vienna> p:abstract ?
                                  abstract .
                                   FILTER langMatches(lang(?abstract), \'de\')
                       }
```

Listing 5.12: A SPARQL query to retrieve a german abstract of search term

```

PREFIX p: <http://dbpedia.org/property/>
             PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
                  SELECT DISTINCT ?reference WHERE {
                      <http://dbpedia.org/resource/Vienna> p:reference ?reference.
             }
```

Listing 5.13: A SPARQL query to retrieve the external links of search term

In order to avoid running the same query into SPARQL endpoint the PEAR Cache_Lite [29] library has been used. Cache_Lite is a little cache system optimized for file containers. The cache_lite uses XML structure to cache the data. Therefore the resultset object retrieved from SPARQL engine must be converted into a valid XML format. In order to solve this problem a method has been implemented that parses the elements from resultset object and converts these into a valid XML form.

5.7.3.2 Geonames module

The geonames module is implemented based on the geografic data using specialized repository (Geonames [30]). It makes use of geonames data stored into a MySQL database containing the names of geografic entities of all over the world in different languages.

As input of this module are the geografic names which are identified in 5.7.2.2 using DBpedia data set. Often geografic names are ambiguous, some of them are homonymous with distinct geografic coordinates pairs of general terms which identify distict places on the world (e.g. "Vienna" in Europe vs. "Vienna" in US) or some temporal geonames (e.g. "Leningrad","Sankt Petersburg").

The identification of geografic names is a difficult task and an ongoing research area. Therefore, this part of work is only focused to identify the geografic populated places that may contain supplemantary information to the user's interest. The exact identification of populated places is done using the four mostly used geoname feature codes [31]:

1. **PPLC** - capital of a political entity (e.g. Vienna in Austria).

2. **PPLA** - seat of a first-order administrative division. PPLC takes precedence over PPLA (e.g. Graz in Austria).

3. **ADM1** - first-order administrative division, a primary administrative division of a country, such as a state in the United States (or Steiermark in Austria).

[29] Cache Lite , http://pear.php.net/package/Cache_Lite/

[30] Geoname, http://www.geonames.org/

[31] Geonames feature codes, http://www.geonames.org/export/codes.html

4. **PPL** - populated place, a city, town, village, or other agglomeration of buildings where people live and work.

Since the name of populated place may exists in different places in the world a method has been implemented which classifies all these places found with specific priorities. There are only two priority low and high. These two priorities are succesively given based on geoname feature codes. The listing 5.14 explains the way of implementation. It retrives the ID of populated place from the geoname table using the geoname feature codes.

Example: Bern (the capital of Switzerland) belongs to the feature code PPLC, which possessed the high priority (pplc = 1), then all other places found with the feature code PPL(e.g. Bern in America/Chicago or Bern in Europe/Amsterdam) have low priority (ppl = 0).

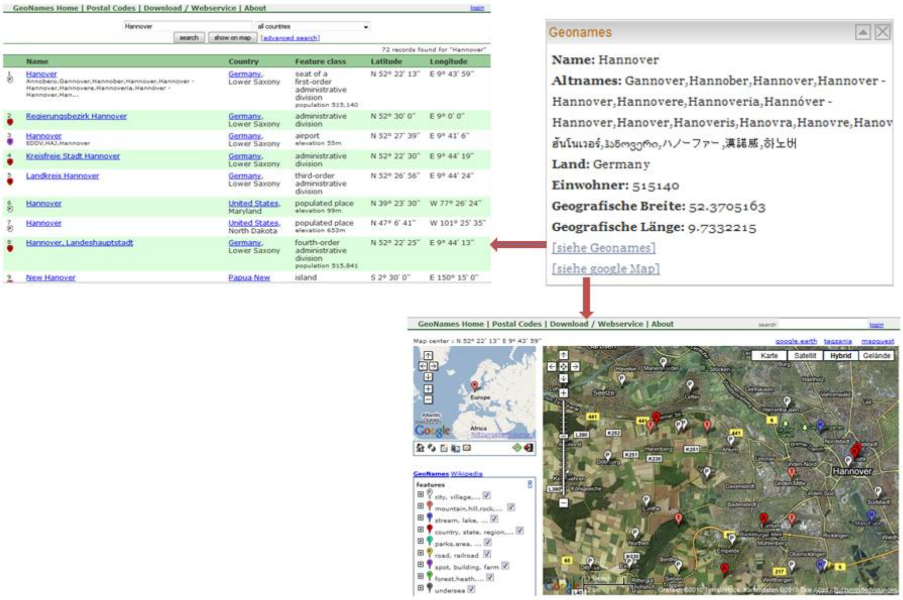
```

function getGeoNameIdWithHighPriority($query){

    $id = 0;
    $pplc = 0;
    $ppla = 0;
    $adm1 = 0;
    $ppl = 0;

    $population_current = 0;
    $result = mysql_query(”SELECT * FROM geoname WHERE name =’$query’”);

    while ($row = mysql_fetch_assoc($result)){

      if($row[’fcode’] == ’PPLC’){
        $id = $row[’geonameid’];
        $pplc = 1;
      }

      else if($row[’fcode’] == ’PPLA’ && $pplc == 0){
        $id = $row[’geonameid’];
        $ppla = 1;
      }


      else if($row[’fcode’] == ’ADM1’ && $pplc == 0 && $ppla == 0){
        $id = $row[’geonameid’];
        $adm1 = 1;
      }

      else if($row[’fcode’] == ’PPL’ && $pplc == 0 && $ppla == 0 && $adm1 == 0 &&
          $row[’population’] > $pop_current){
        $id = $row[’geonameid’];
        $population_current = $row[’population’];
        $ppl = 1;
      }
      else if ($pplc == 0 && $ppla == 0 && $adm1 == 0 && $ppl == 0) {
        $id = $row[’geonameid’];
      }

    }
    return $id;
  }
```

Listing 5.14: Retrieving of ID from geoname tables using the geoname feature codes

Once the places are identified it is possible to retrieve further geografic information from geonames database. As we can see in figure 5.16 the term named "Hannover" has been identified as a populated place which contain the alternative names, country, population, geografic coordinates (latitude and Longitude) and the two external links.

The first external links "siehe Geonames" use the GeoNames URL (http://www.geonames.org/search.html?q=Graz where q represent the query. This link redirects users to the result of search query. The result contain all geografic information that deal with this query without specifying the exact match.

The second link "siehe google map" redirects users to the google map being modified by GeoNames. This displays at center the of the map the place queried with geografic coordinates retrieved from the local database of geoname. Around this place are other neighbors or places with the distance against it.



GeoNames Home | Postal Codes | Download / Webservice | About
login
Hannover
all countries
search
show on map
[advanced search]

72 records found for "Hannover"

	Name	Country	Feature class	Latitude	Longitude
1	Hanover Annobero,Gannover,Hannober,Hannover,Hannover - Hannover,Hannovere,Hannoveria,Hannóver - Hannover,Han...	Germany, Lower Saxony	seat of a first-order administrative division population 515,140	N 52° 22' 13"	E 9° 43' 59"
2	Regierungsbezirk Hannover	Germany, Lower Saxony	administrative division	N 52° 30' 0"	E 9° 0' 0"
3	Hannover EDDV,HAJ,Hannover	Germany, Lower Saxony	airport elevation 55m	N 52° 27' 39"	E 9° 41' 6"
4	Kreisfreie Stadt Hannover	Germany, Lower Saxony	administrative division	N 52° 22' 30"	E 9° 44' 19"
5	Landkreis Hannover	Germany, Lower Saxony	third-order administrative division	N 52° 26' 56"	E 9° 44' 24"
6	Hannover	United States, Maryland	populated place elevation 99m	N 39° 23' 30"	W 77° 26' 24"
7	Hannover	United States, North Dakota	populated place elevation 653m	N 47° 6' 41"	W 101° 25' 35"
8	Hannover, Landeshauptstadt	Germany, Lower Saxony	fourth-order administrative division population 515,841	N 52° 22' 25"	E 9° 44' 13"
9	New Hanover	Papua New	island	S 2° 30' 0"	E 150° 15' 0"

Geonames

Name: Hannover
Altnames: Gannover,Hannober,Hannover,Hannover - Hannover,Hannovere,Hannoveria,Hannóver - Hannover,Hanover,Hanoveris,Hanovra,Hanovre,Hanov
ฮันโนเวอร์,ჰანოვერი,ハノーファー,漢諾威,하노버
Land: Germany
Einwohner: 515140
Geografische Breite: 52.3705163
Geografische Länge: 9.7332215
[siehe Geonames]
[siehe google Map]

Figure 5.16: Term-Related Information from Geonames

5.7.3.3 Flickr module

This module was implemented using Flickr API. Flickr API provides different methods [32] on photo viewing, manipulating and search. All these methods are invoked using REST (see 5.5.5.4) protocols. The listing 5.15 represents a part of code which creates the flickr REST URL.

In order to create a flickr REST URL the following parameters must be set:

1. **method**, is a Flickr API method. E.g. flickr.photos.search returns a list of photos matching some criteria.

2. **api_key**, is an API application key.

3. **text**, is a search term extracted from the text. Photos who's title, description or tags contain the term will be returned.

[32] Flickr Methods,http://www.flickr.com/services/api/

4. **per_page**, is the number of photos to return per page. The value set is 16.
5. **sort**, is an order in which to sort returned photos. The possible values are: date-posted-asc, date-posted-desc, date-taken-asc, date-taken-desc, interestingness-desc, interestingness-asc, and relevance.
6. **lat and lon**, are valid latitude and longitude, in decimal format, for doing radial geo queries. These parameters are only set if the term extacted is a toponym.
7. **radius**, is a valid radius used for geo queries, greater than zero and less than 20 miles (or 32 kilometers), for use with point-based geo queries. The default value is 5 (km).

All these parameters are appended to the base REST URL (http://www.flickr.com/services/rest/) which returns the result in XML form using php method *file_get_contents*. The result retrieved includes a set of photo attributes corresponding to each term extracted from the text. The set of photo attributes are parsed using Mini-XML. Mini-XML [33] is a small XML library that you can use to read XML. Same as by DBpedia Module (see 5.7.3.1) an cache system (PEAR Cache_Lite) has been implemented to store information that has been once retrieved.

The listing 5.16 explains the way of parsing of result retrieved from flickr API. The final result include the path of flickr images, titles of images and URLs of flickr images. The module integrated in SemTex application is illustrated in Figure 5.17. Each photo contain a title (visible with mouse-over) and a link directing to Flickr page where the photo has been uploaded.

```

    // Create a flickr rest url. Update the method and various parameters
    // below to match the request you want to make to the API.
    $base = 'http://www.flickr.com/services/rest/';
    $query_string = '';
    $params = array(
      'method'  => 'flickr.photos.search',
      'api_key'  => '6537e4e69ff7cf2761661a712b5abb53',
      'text'  =>  $topic,
      'per_page'      => '16',
      'sort'         => 'relevance'
    );

    if(!is_null($lat) && !is_null($long) && !is_null($radiusKm)){
      $params['lat'] = $lat;
      $params['lon'] = $lon;
      $params['radius'] = $radiusKm;
    }


    foreach ($params as $key => $value) {
      $query_string .= "$key=" . urlencode($value) . "&";
    }
    $url = "$base?$query_string";
```

Listing 5.15: Create a flickr rest url

5.7.3.4 YouTube module

This module retrieves video information using YouTube Data API. The YouTube Data API allows client applications to retrieve and update YouTube content in the form of Google Data API feeds. The SemTex application uses the YouTube Data API only to query for videos that match particular criteria and does not require any authentication.

The YouTube Data API has been accessed through the use of PHP libraries so called Zend_Gdata (see 5.5.5.2) which is distributed as part of the Zend Framework [34].

[33] Mini-XML, http://www.minixml.org/

[34] Zend Framework, http://framework.zend.com/

```

if ( $xml ) {

  // Parse the xml using minixml.
  $parsed = new MiniXMLDoc();
  $parsed->fromString($xml);
  $root =& $parsed->getRoot();

  // Get the xml rsp element and stat attribute.
  $rsp =& $root->getElement(’rsp’);
  $stat =  $rsp->attribute(’stat’);

  // Check that the api request was succesful.
  if ($stat != "fail") {

    /*
     In this example we’re using the search method but you would
     have different elements and attrributes depending on the
     method you called above. Check the example xml responces
     in the flickr api documentation for what xml elements and
     atributes you can expect to see returned.
     */

    // Get the "photo" xml element.
    $photos =& $root->getElement(’photos’);
    $children =& $photos->getAllChildren(’photo’);
    // Check the "photo" element was found.

    $numChildren = $photos->numChildren(’photo’);

    if ($photos) {

      for($i= 0; $numChildren > $i; $i++){


        // Load the "photo" attributes into vars for later use.
        $photoid =  $children[$i]->attribute(’id’);
        $secret =  $children[$i]->attribute(’secret’);
        $server =  $children[$i]->attribute(’server’);
        $title =  $children[$i]->attribute(’title’);
        $owner =  $children[$i]->attribute(’owner’);

        // Turn the photo attributes into a flickr image url.
        $photo = "http://www.flickr.com/photos/" . $photoid . "_" . $secret . "_m.jpg";
        $photo_flickr = "http://www.flickr.com/photos/" . $owner . "/".$photoid."/";
        // The result! Show the latest sunset photo.
        $photos_array[$i][’url’] = $photo;
        $photos_array[$i][’title’] = $title;
        $photos_array[$i][’flickr’] = $photo_flickr;

      }
    } else {
      echo("The photo element was not found.");
    }
  } else {
    echo("Requested failed. Check the method and parameters sent.");
  }
}
```

Listing 5.16: Parsing of Flickr’s XML result

The listing 5.17 explains in details the use of Zend_Gdata in fetching YouTube data which is implemented in our application.

Because we are fetching only data, we are not required at this point to use any parameters for authentication. So the *Zend_Gdata_Youtube()* constructor remains empty. Ordinarily, the constructor accepts four parameters, however: a Zend_HTTP object, application ID, client ID, and developer key. After the $yt object is created, a new query will be created using the *newVideoQuery()* method. The query allows users to set any number of properties. In this case, set the search terms (extracted from our text) from along with starting index and total number of results we want to return. After a query was created, the query is passed into the *getVideoFeed()* method, which returns a Zend_Gdata_YouTube_VideoFeed object. Next, the Zend_Gdata_YouTube_VideoFeed was taken and looped through each video in the object. Because each video in the object is a Zend_Gdata_YouTube object, we use the the following

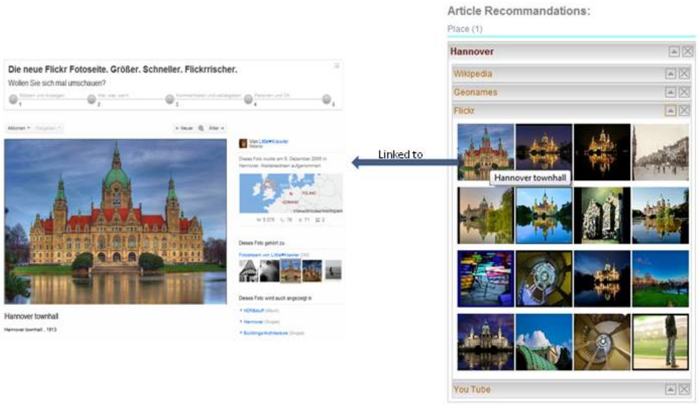


Figure 5.17: Term-Related Photos from Flickr

```
$yt = new Zend_Gdata_YouTube();
  $query = $yt->newVideoQuery();
  $query->setQuery($searchTerm);
  $query->setStartIndex($startIndex);
  $query->setMaxResults($maxResults);

  // results from youtube
  $feed = $yt->getVideoFeed($query);

    foreach ($feed as $key => $entry) {

      $thumbnails = $entry->getVideoThumbnails();
      $thumbnailUrl = $thumbnails[1][’url’];
      $ytUrl = $entry->getVideoWatchPageUrl();
      $videoTitle = $entry->getVideoTitle();
      $videoDuration = $entry->getVideoDuration();
      $videoViewCount = $entry->getVideoViewCount();

      // results stored in an array variable
      $videos_array[$key][’thumbnailUrl’] = $thumbnailUrl;
      $videos_array[$key][’ytUrl’] = $ytUrl;
      $videos_array[$key][’videoTitle’] = $videoTitle;
      $videos_array[$key][’videoDuration’] = $videoDuration;
      $videos_array[$key][’videoViewCount’] = $videoViewCount;

    }
```

Listing 5.17: Fetching YouTube data using Zend_Gdata library

methods such as getVideoThumbnails(), getVideoWatchPageUrl(), getVideoTitle(), getVideoDuration() and getVideoViewCount(). These methods are used to built an array result which will be later used to display the video information gained from YouTube to the users or editors. The Figure 5.18 shows the video information gained from YouTube based on keyword search ”Hannover”.

Additionaly, an cache system (PEAR Cache_Lite) has been implemented to store information that has been once retrieved.

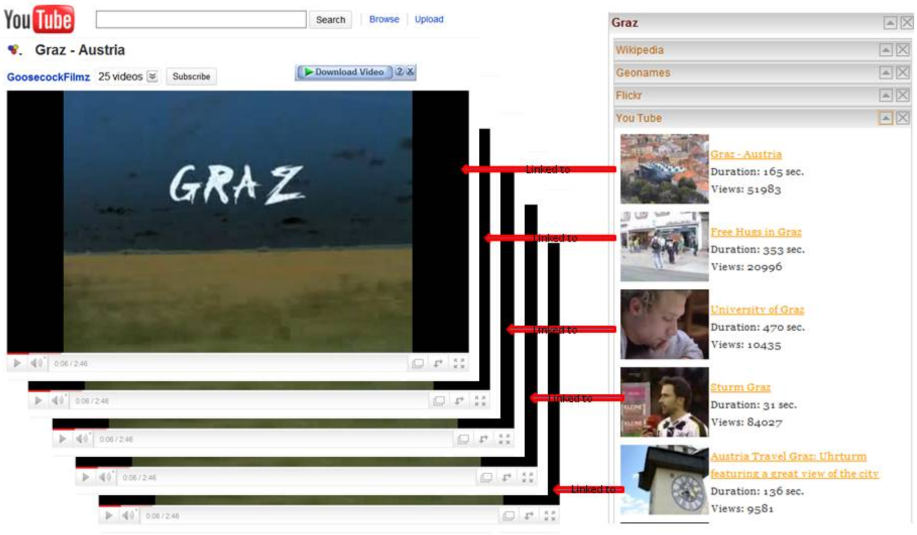


Figure 5.18: Term-Related Videos from YouTube

5.7.4 Data Publication

The previous sections described implemented modules of consuming data from Linked Data. This section describes how to publish information conform to Linked Data principles using formats that allows machines to access and re-use the information such as RDFa(see section 3.3.1) or placing a static RDF/XML document on web server. There are only two ways of accessing this information; using a standard SPARQL query interface(see SPARQL Endpoint) or using indexing machines like Sindice [35].

In following two main methods of data publication module will be presented; a method for creating a XHTML page with embedded RDF and a method for extracting RDF from XHTML.

5.7.4.1 Creating a XHTML page with embedded RDF

Embedding of RDFa attributes in the published paqes of SemTex application plays an important role for consuming of our information from the third party. It means that the related information will be picked up automatically by third party. In order to publish our user generated information conform Linked Data principles the following steps has been taken:

1. DOCTYPE and Validation

2. Adding Vocabularies and

3. Marking-up Information

DOCTYPE and Validation

DOCTYPE is added at the top of XHTML+RDFa page as it shown in listing 5.18.

[35] Sindice, http://sindice.com/

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" "http://www.w3.org/MarkUp/DTD/xhtml-
    rdfa-1.dtd">
```

Listing 5.18: DOCTYPE and Validation

It will be used to validate the XHTML+RDFa page to ensure they are picked up by third party. W3C Validator [36] is the standard validator to check the markup validity of Web documents in HTML, XHTML, SMIL, MathML, etc. But also RDFa Distiller [37] is a useful tool for developing and checking RDFa markup.

Adding RDF Vocabularies

RDF Vocablaries are defined as a set of related classes and properties that can be used in an application for describing the structure and semantics of the web page. The 5.2 shows the RDF vocabularies used in published information of SemTex .

Table 5.2: Vocabularies used in SemTex

List names	Description	Prefix
DBPedia	A cross-domain ontology to describe structured information from Wikipedia	dbpprop
Dublin Core	A common list of terms which are applicable in a wide range of situations.	dcterms
FOAF	The Friend-of-a-Friend vocabulary, used to describe people, companies, projects, and so on	foaf
Geonames	A geographical Vocabulary, used to describe places	geo
SearchMonkey Media	A media Vocabulary, used to describe information on various media types, features, specifications, and items	media
Open Graph Protocol [38]	It lets web pages be rich object in a social graph	og

In the xHTML+RDFa page each RDF vocabulary is identified by the namespace URI. RDFa provides a mechanism to use Compact URIs or CURIEs which involves using a prefix to replace part of the URI. In our SemTex application these prefixes (see Listing 5.19) are declared to the root of the XHTML+RDFa page.

```
<html xmlns="http://www.w3.org/1999/xhtml"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dbpprop="http://dbpedia.org/property/"
  xmlns:dcterms="http://purl.org/dc/elements/1.1/"
  xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
  xmlns:foaf="http://xmlns.com/foaf/01/"
  xmlns:media="http://search.yahoo.com/searchmonkey/media/">
  xmlns:og="http://ogp.me/ns#

...
</html>
```

Listing 5.19: Prefix mapping

Marking-up Information

[36] W3C Validator, http://validator.w3.org/

[37] RDFa Distiller, http://www.w3.org/2007/08/pyRdfa/

After declaring all required namespaces in the page head, other information will be marked with RDFa attributes in the document. Everything that will be marked with an RDFa attributes must have an URI so that others can refer to our data. This is the key principle of Linked Data. In order to let our content te be identified by the others we have to turn our web page into graph objects. Therefore we have used the Open Graph Protocol. Based on the initial state of the protocol on RDFa we place additional meta tags in the head of our web page as it shown in the listing 5.20).

```
<html
...
<head>
...
  <meta property="og:url" content="http://localhost/SemTex/Publisher/Data/Cyberangriff auf
      Wikileaks-Gegner.html" />
  <meta property="og:title" content="Cyberangriff auf Wikileaks-Gegner" />
  <meta property="og:type" content="article" />
  ...
</head>
...
</html>
```

Listing 5.20: Metadata to identify our web page

The first property identifies the url of Web page. Then the second property defines the title of our article. And the last property is used to defines the type of our object, in our case it is *article*. There are also two other properties defined in Open Graph Protocol[39] which are optional; og:description and og:site_name.

There are also other information to be marked-up with RDFa attributes especially the extracted keywords or the related concepts from Linked Data cloud. All these keywords are related to generated article. So we need to establish a connection between the article and the keywords, which we do with the Dublin Core relation property.

```
<span resource = "http://dbpedia.org/resource/Graz" rel = "dcterms:relation">
 <span xml:lang="de" property="rdfs:label">Graz</span>
 <span xml:lang="de" property="dbpprop:abstract" content = "Graz ist die Landeshauptstadt
     der Steiermark und mit 294.000 Einwohnern ..."></span>
 <span property="dbpprop:reference" content = "http://www.graz.at"></span>
 ...
</span>
```

Listing 5.21: Adding RDFa attributes to the information from DBPedia

The Listing 5.21 shows a RDFa-enhanced XHTML snippet of DBPedia resource which is related to the keyword ”Graz”. It includes name, short description and the external links from Wikipedia about the keyword ”Graz”. We use the property label from RDF schema to indicate the name. The last two DBPedia properties indicate the abstract and the external links from Wikipedia. A special attribute named xml:lang specifies the language used in the contents.

```
<span resource = "http://www.geonames.org/2778067/" rel = "dcterms:relation">
 <span property="geo:lat" content = "47.0666667"></span>
 <span property="geo:long" content = "15.4500000"></span>
</span>
```

Listing 5.22: Adding RDFa attributes to the information from Geonames

[39] Open Graph Protocol, http://ogp.me/

The listing 5.22 indicates some other related geographical information from Geonames which means if the keyword is a place. It includes the lat(itude) and long(itude) which are specified from the geoname ontology.

```
<span resource = "http://www.flickr.com/photos/64897154@N00/2289135129/" rel = "dcterms:
    relation">
 <span xml:lang="de" property="dcterms:title" content = "Pristina"></span>
 <span rel="foaf:depiction" resource = "http://www.flickr.com/photos/2289135129_e183e6eba3_m
     .jpg"></span>
</span>
```

Listing 5.23: Adding RDFa attributes to the information from Flickr

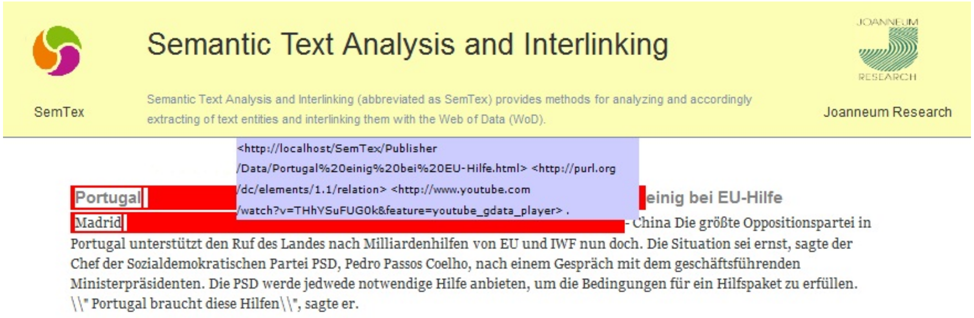
The listing 5.23 shows the markups of photos from Flickr, which are also related. The attribute foaf:depiction represents the according FOAF property which is used to specify Flickr photos.

```
<span resource = "http://www.youtube.com/watch?v=5jjORffaPCo&feature=youtube_gdata_player"
    rel = "dcterms:relation">
 <span xml:lang="de" property="dcterms:title" content = "Faith No More - Pristina"></span>
 <span rel="media:Video" resource = "http://www.youtube.com/watch?v=5jjORffaPCo&feature=
     youtube_gdata_player"></span>
 <span rel="media:Thumbnail" resource = "http://i.ytimg.com/vi/5jjORffaPCo/1.jpg"></span>
 <span property="media:duration" content = "PT312S"></span>
 <span property="media:views" content = "1235"></span>
</span>
```

Listing 5.24: Adding RDFa attributes to the information from Youtube

The listing 5.24 indicates the youtube or the video information using SearchMonkey media vocabulary. The properties media:Video and media:Thumbnail specify the sources in which the keyword or the user generated article is related. The properties media:duration and media:views indicate video duration using XML Schema duration format and the number of times a video has been viewed.

As we can see from listings above the content which is included from other sources is not human-visible. The content provides links or labels that are machine-readable. In this way machines or specialized agents from other platforms can retrieve other related sources from user generated content [Adida and Birbeck, 2008]. To test the added RDFa annotations the RDFa Highlight bookmarklet has been used. The figure 5.19 shows the highlighed terms which provide RDF informations. Next section will explain how to generate a RDF document from the document beeing published.



Semantic Text Analysis and Interlinking

Semantic Text Analysis and Interlinking (abbreviated as SemTex) provides methods for analyzing and accordingly extracting of text entities and interlinking them with the Web of Data (WoD).

<http://localhost/SemTex/Publisher/Data/Portugal%20einig%20bei%20EU-Hilfe.html> <http://purl.org/dc/elements/1.1/relation> <http://www.youtube.com/watch?v=THhYSuFUG0k&feature=youtube_gdata_player> .

Portugal einig bei EU-Hilfe

Madrid - China Die größte Oppositionspartei in Portugal unterstützt den Ruf des Landes nach Milliardenhilfen von EU und IWF nun doch. Die Situation sei ernst, sagte der Chef der Sozialdemokratischen Partei PSD, Pedro Passos Coelho, nach einem Gespräch mit dem geschäftsführenden Ministerpräsidenten. Die PSD werde jedwede notwendige Hilfe anbieten, um die Bedingungen für ein Hilfspaket zu erfüllen. \\" Portugal braucht diese Hilfen\\", sagte er.

Figure 5.19: RDFa Highlight

5.7.4.2 Extracting RDF from XHTML

As we have seen in the previous section that RDFa enables to integrate RDF in encoded XHTML. This ability allows to easily extract a set of RDF triples from an RDFa annotated XHTML using ARC from the section 5.5.5.1. ARC provides dedicated parsers for a specific serialization (see section 3.2.1.2) such as RDF/XML or Turtle. In the listing 5.25 is an PHP code snippet of Data publication module which uses ARC. ARC reads and parses html document into a tree structure only once, and then a dedicated extractor is applied to extract RDF triples from the tree. After extraction process the RDF triples array will be serialized as XML which will be stored both localy as RDF/XML file and in the ARC Store. The structure of RDF/XML document is shown in the listing 5.26.

```
$page = "http://localhost/Semtex/Data/Article.html";
$ns = array(
  ’dcterms’=>’http://purl.org/dc/elements/1.1/’,
  ’rdfs’=>’http://www.w3.org/2000/01/rdf-schema#’,
  ’geo’=>’http://www.w3.org/2003/01/geo/wgs84_pos#’,
  ’media’=>’http://search.yahoo.com/searchmonkey/media/’,
  ’foaf’=>’http://xmlns.com/foaf/01/’,
  ’og’=>’http://ogp.me/ns#’,
  ’dbpprop’=>’http://dbpedia.org/ontology/’
);

$config = array(’auto_extract’ => 0);
$parser = ARC2::getSemHTMLParser();
$parser->parse($page);
$parser->extractRDF(’rdfa’);

$triples = $parser->getTriples();
$rdfxml = $parser->toRDFXML($triples,$ns);
```

Listing 5.25: Using ARC to extract and serialize RDF

5.7.5 SemTex UI and Visualization

From online content providers we learned that professional online editors spend most of their time investigating interesting and suitable material, which may enrich their editorial content to become more fascinating and valuable to their audience and/or enables them to produce editorial content quicker, enriched with multimedia objects and hyperlinks to advanced material. At this point this thesis was aimed to develop SemTex that is very efficient (i.e. developing their editorial content and getting it published instantly on the Web) and providing up to date (multimedia) content from Linked Open Data, particularly capable of drawing the attention of readers.

Therefore, SemTex provides a simple user interaction running on server as script based Web application. So that the user does not have to worry about installing any prerequisites on the target system. With a simple click on the button *Web Search* the editor recieves all the related information from the Web of Data about the article edited.

The YUI from section 5.5.5.7 as development library aiming at implementation of Web applications with set of utilities, controls written in JavaScript and CSS library has been approved as scalable, fast, and robust for the realization of the user interface. The YUI uses techniques such as DOM scripting, DHTML and AJAX.

The design of user interface includes three views which are illustrated in figure **??** . The first view on the left side is the text editor area where the original text is inserted by an editor. The second view or the panel on the left side contains the results from Web of Data. And the last view highlights the results of extracted terms from the text.

The text editor area is an user interface control that allows for the rich formatting of text content, including common structural text like lists, formatting part of text like bold, italic and underline text, and

```

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/elements/1.1/"
  xmlns:profile="http://www.w3.org/1999/xhtml/vocab#"
  xmlns:og="http://ogp.me/ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dbpprop="http://dbpedia.org/ontology/"
  xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
  xmlns:foaf="http://xmlns.com/foaf/01/"
  xmlns:media="http://search.yahoo.com/searchmonkey/media/">

  <rdf:Description rdf:about="http://localhost/SemTex/Publisher/Data/Article.html">
    <dcterms:title>Article</dcterms:title>
    <dcterms:title xml:lang="de">Article</dcterms:title>
    <og:url xml:lang="de">http://localhost/SemTex/Publisher/Data/Article.html</og:url>
    <dcterms:type xml:lang="de">article</dcterms:type>
    <og:title xml:lang="de">Article</og:title>
    <dcterms:relation rdf:resource="http://dbpedia.org/resource/Graz"/>
    <dcterms:relation rdf:resource="http://www.geonames.org/2778067/"/>
    <dcterms:relation rdf:resource="http://www.flickr.com/photos/13290612@N04/5455810453/"/>
    <dcterms:relation rdf:resource="http://www.youtube.com/watch?v=ZDrIBxShUCw&amp;feature=
        youtube_gdata_player"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://dbpedia.org/resource/Graz">
    <rdfs:label xml:lang="de">Graz</rdfs:label>
    <dbpprop:abstract xml:lang="de">Graz ist die Landeshauptstadt der Steiermark und mit
        290.316 ..</dbpprop:abstract>
  </rdf:Description>

  <rdf:Description rdf:about="http://www.flickr.com/photos/13290612@N04/5455810453/">
    <dcterms:title xml:lang="de">Natural light over Graz</dcterms:title>
    <foaf:depiction rdf:resource="http://www.flickr.com/photos/5455810453_33918ce62a_m.jpg
        "/>
  </rdf:Description>

  <rdf:Description rdf:about="http://www.youtube.com/watch?v=ZDrIBxShUCw&amp;feature=
      youtube_gdata_player">
    <dcterms:title xml:lang="de">Graz - Austria</dcterms:title>
    <media:Video rdf:resource="http://www.youtube.com/watch?v=ZDrIBxShUCw&amp;feature=
        youtube_gdata_player"/>
    <media:Thumbnail rdf:resource="http://i.ytimg.com/vi/ZDrIBxShUCw/1.jpg"/>
    <media:duration xml:lang="de">165</media:duration>
    <media:views xml:lang="de">61560</media:views>
  </rdf:Description>

</rdf:RDF>
```

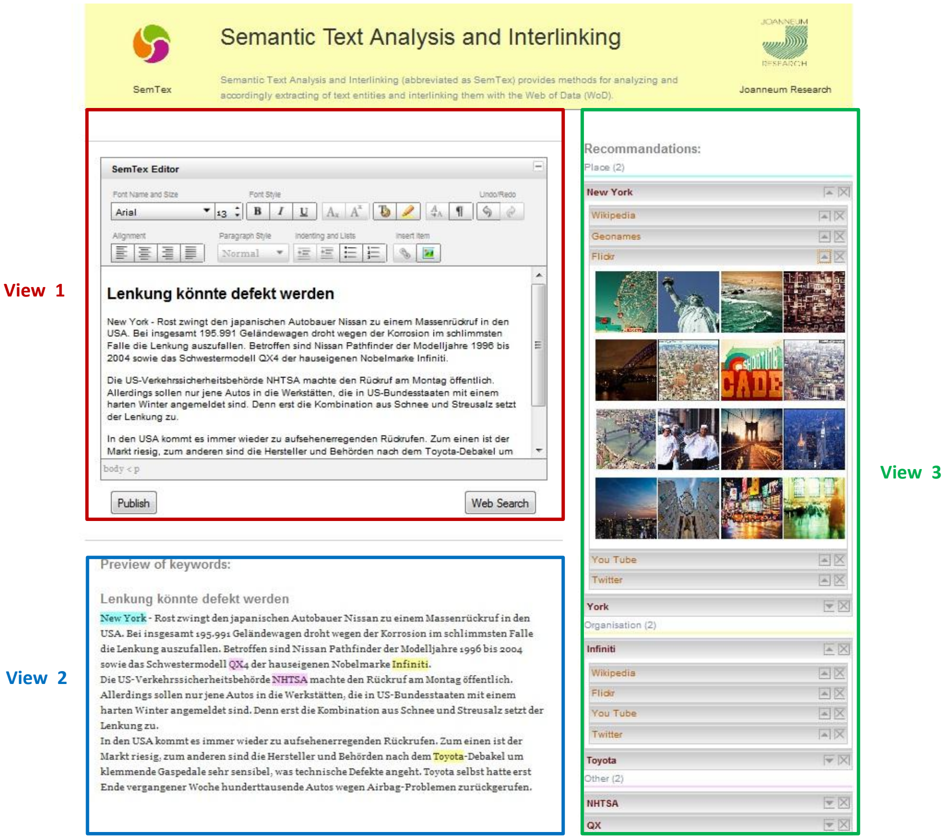
Listing 5.26: RDF/XML structure of a published page

drag-and-drop inclusion, inserting and sizing of images. The text editor area is an YUI Simple Editor Control that is used to fulfill the editor needs to write an article.

The second view, in this case the data requested from Linked Data are visualized using Javascript Bubbling Library [40]. Javascript Bubbling Library is a set of plugins and widgets, for building event-driven web applications using the bubble-up technique. The Bubbling Library includes several plugins that can be used to extend the YUI Library to manage dynamic areas. YUI Widget: Nested Accordions Menus has been used to view the information from various sources from Linked Open Data about the term extracted from the text. Each term contains 5 sources including Wikipedia, Geonames, Flickr, You Tube and Twitter. And each of these sources contain their own information related to the term.

The *publish* button on the right side underneath the textarea is then activated by an editor when the editor has completely written the article. The article will be integrated in the final online news article and RDF describing (see section 5.7.4) the news article as well as links to other resources is also supplied.

[40] Bubbling Library, http://www.bubbling-library.com/



Semantic Text Analysis and Interlinking (abbreviated as SemTex) provides methods for analyzing and accordingly extracting of text entities and interlinking them with the Web of Data (WoD).

Lenkung könnte defekt werden

New York - Rost zwingt den japanischen Autobauer Nissan zu einem Massenrückruf in den USA. Bei insgesamt 195.991 Geländewagen droht wegen der Korrosion im schlimmsten Falle die Lenkung auszufallen. Betroffen sind Nissan Pathfinder der Modelljahre 1996 bis 2004 sowie das Schwestermodell QX4 der hauseigenen Nobelmarke Infiniti.

Die US-Verkehrssicherheitsbehörde NHTSA machte den Rückruf am Montag öffentlich. Allerdings sollen nur jene Autos in die Werkstätten, die in US-Bundesstaaten mit einem harten Winter angemeldet sind. Denn erst die Kombination aus Schnee und Streusalz setzt der Lenkung zu.

In den USA kommt es immer wieder zu aufsehenerregenden Rückrufen. Zum einen ist der Markt riesig, zum anderen sind die Hersteller und Behörden nach dem Toyota-Debakel um

Preview of keywords:

Lenkung könnte defekt werden

New York - Rost zwingt den japanischen Autobauer Nissan zu einem Massenrückruf in den USA. Bei insgesamt 195.991 Geländewagen droht wegen der Korrosion im schlimmsten Falle die Lenkung auszufallen. Betroffen sind Nissan Pathfinder der Modelljahre 1996 bis 2004 sowie das Schwestermodell QX4 der hauseigenen Nobelmarke Infiniti.
Die US-Verkehrssicherheitsbehörde NHTSA machte den Rückruf am Montag öffentlich. Allerdings sollen nur jene Autos in die Werkstätten, die in US-Bundesstaaten mit einem harten Winter angemeldet sind. Denn erst die Kombination aus Schnee und Streusalz setzt der Lenkung zu.
In den USA kommt es immer wieder zu aufsehenerregenden Rückrufen. Zum einen ist der Markt riesig, zum anderen sind die Hersteller und Behörden nach dem Toyota-Debakel um klemmende Gaspedale sehr sensibel, was technische Defekte angeht. Toyota selbst hatte erst Ende vergangener Woche hunderttausende Autos wegen Airbag-Problemen zurückgerufen.

Figure 5.20: SemTex UI layout

5.8 Evaluation

Although the goal of thesis was to investigate paradigms of the commercial adoption of Linked Data, and even though the SemTex is still under development, so the preliminary evaluation has been conducted and included a benchmark against other systems for semantic content enrichement. These early tests already indicate that the achieved performance of SemTex prototype is superior compared to similar systems that aim at enhancing (editorial) content. The most distinguishing features of our solution are the support of the German language, recommendation of multimedia content and the relationship of the recommended objects to concepts extracted from the text. One goal of this evalation was to benchmark our prototype against three popular applications for semantic enrichment as listed in table 5.3 where demonstrators are publicly available on the Web: OpenCalais, Ontos Semantic API, and Zemanta. However, as can be seen from the comparison table, some tools have restrictions which resulted in a limited evaluation: We found out that Calais currently does not support German-language content at all. Using a translation service as an intermediary step and processing the translated output an English version of the German-language article - might at the first glance solve the language problem. As this would indirectly result in an evaluation of the translation service within our evaluation, we decided not to include OpenCalais at all as its functionality does not match with our requirements. We furthermore learned that Ontos Semantic API is currently in fact capable of extracting terms, but does not provide any content enrichment which resulted in a drop-out of Ontos Semantic API, too. As only Zemanta provides all required functionality for a valid comparison, we are limited to benchmark our prototype with Zemanta.

The following six-step procedure for this preliminary evaluation has been developed:

1. In the first step a set of German-language articles has been selected from a local commercial news provider. The preliminary evaluation contained only ten articles, including articles in the domains of politics, business, sports, and culture. For a more comprehensive future evaluation we have access to more than 3,000 articles from the news paper.

2. Manual extraction of persons, places, and organizations from each article. Furthermore, extraction of terms which are important for a particular article as they describe the content.

3. Thereafter, a comparison of content processing of each article with SemTex as well as with Zemanta. This resulted in an automatic extraction of terms which we evaluate in the next step.

4. A quantitative comparison between manually and automatically extracted terms for both applications, SemTex and Zemanta.

5. A manual analysis of quantity and quality of the content recommended from both applications. Thereby, the intention was to explore whether and how recommended objects relate to the terms (concepts) extracted from the articles.

6. A comparison of recommendation-ability of our prototype to the results of Zemanta. As the ability to suggest content for enrichement differs in various aspects, this evaluation is mainly done qualitatively.

In the following, I present the lessons learned from this preliminary evaluation. Figures 5.21 and 5.22 present the quantitative results of the term extraction evaluation (step 1 - 4): From the investigation, we came to the conclusion that the SemTex prototype has generated satisfactorily results. The recall of all relevant terms (cf. Figure 4) is considerably higher in SemTex prototype and even though Zemanta has identified less relevant terms the precision values are comparable. This shows that the SemTex prototype is capable of retrieving more correct relevant terms which allows online editors to be more flexible when enriching their content with content from third party sources. Figure 5 shows the recall values for places, people, and organizations. Our prototype outperforms Zemanta when extracting people and places: In this context it is important to mention that Zemanta can extract internationally known persons very well, but struggles when trying to extract local politicians or businessmen. When it comes to extracting organizations, this prototype is on a comparable level with Zemanta.

Evaluating the ability of the SemTex prototype to suggest text, pictures and videos for content enrichement (step 5 - 6) is more challgenging, especially as it differs from those of Zemanta in various aspects: While SemTex links suitable data for any of the selected terms using Wikipedia, Geonames, Flickr, Twitter and Youtube, Zemanta only recommends content for the entire article. Therefore, data recommended by Semtex is of finer granularity and of higher semantics. This aspect makes it almost impossible to compare the quality of content linked by the SemTex against the content Zemanta recommends. From the requirements of online editors, we learned that content should preferably be linked to the terms (concepts) in the text and not to the entire article. Evaluating the quantity of content linked, we found out that the SemTex prototype suggests a fullness of media and text objects compared to Zemanta. However, both quantity and quality of content can be further improved: As being noticed, that mainly pictures from Flickr and videos from Youtube did not always match with extracted concepts.

Table 5.3: Comparison of different systems for semantic enrichment

	English language support	German language support	Categorizations	Integration of external content
SemTex	via extensions	yes	yes	yes (Linked Data, multimedia content)
OpenCalais	yes	-	yes	-
Zemanta	yes	limited	-	yes (select sources)
Ontos Semantic API	yes	yes	yes	-

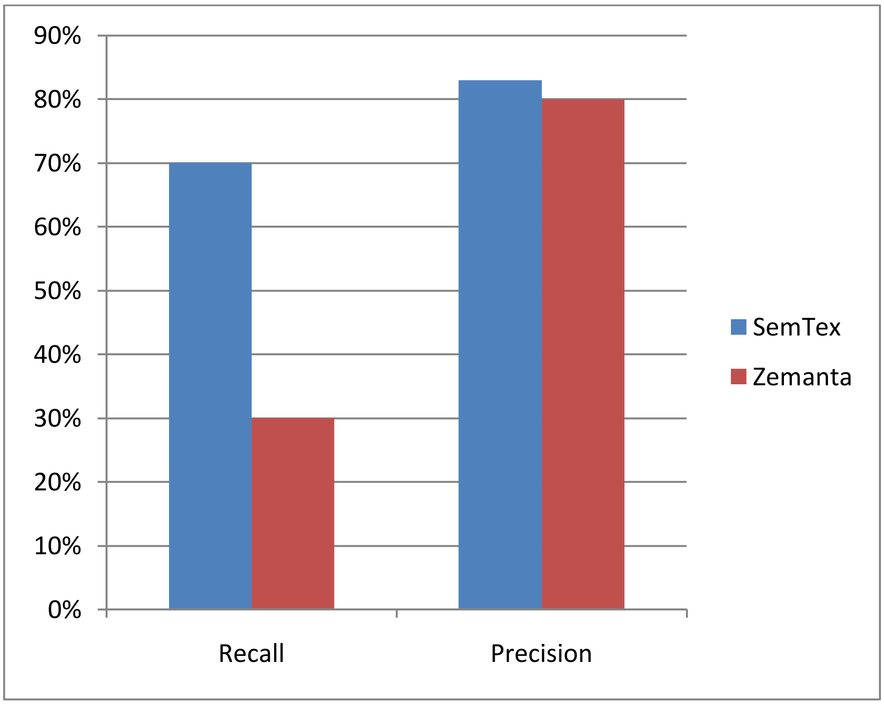


Figure 5.21: Recall and precision for all terms

5.9 Challenges

Over the past months Linked Data technology has matured, various applications have been created, and new datails constantly being added. However, during the development of the prototype we have identifed three important challenges that still need attention and should be addressed to make Linked Data applications competitive and allow their use in professional production environments. It has to be noted that most of the issues do not only apply to Linked Data but the Web in general.

Distributed infrastructure: One of the advantages of Linked Data is that it can be accessed via HTTP URIs and queries on remote SPARQL [Prud'hommeaux and Seaborne, 2008] endpoints that are made available by the data providers themselves or Linked Data consolidators (such as the Virtuoso instance hosting Linked Open Data [41]). Even though it is appealing to use data from the Web without having to care about own infrastructure the downside is that one becomes dependent on the data providers' infrastructure. Response times for queries depend on server load and it could happen that a resource is unavailable for any (technical) reason that cannot be infuenced by the data consumer. One of the solutions to this issue is hosting a copy of relevant Linked Data on infrastructure that is under one's own control - which in turn creates the need for getting informed about dataset changes where

[41] LOD http://lod.openlinksw.com/

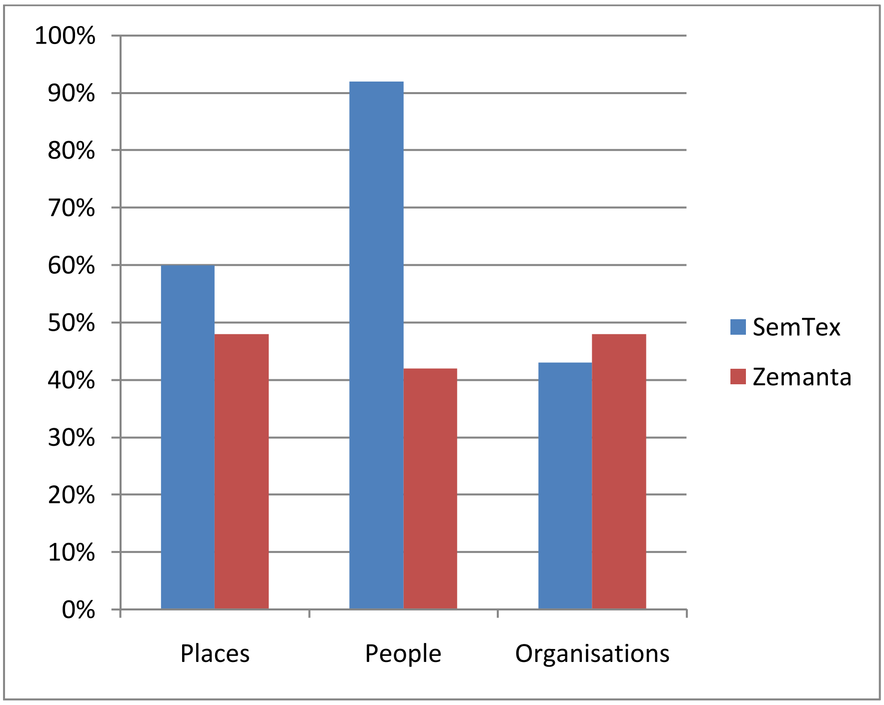


Figure 5.22: Recall of entities by category

different approaches exist and a commonly agreed strategy still needs to evolve [Lyons et al., 2009]. From a business perspective one could also identify the need for Linked Data providers/consolidators that offer Service Level Agreements that contractually guarantee the availability of relevant Linked Data for professional users.

Data quality: For professional users the quality of data can be an important issue that should be taken into account by data publishers if they want their data to be reused.

Legal issues: Dataset publishers should be explicit about licensing terms of their data to protect their rights or make the data legally safely useable by others [Davis, July 2009]. Even though further challenges in the context of Linked Data (such as provenance, trust, archiving, etc.) would be worthwhile dealing with it seems that solving the issues presented above could foster an even greater acceptance of solutions based on Linked Data for professional use.

5.10 Summary

In this chapter the details referring to the Semtex aplication including introduction, scenarios using Linked Data, the system architectural design, use cases, implementation, evaluation and challenges were introduced. This chapter explained much about professional production of valuable and up-to-date online content as outlined in section 5.1. The intention of this work was to to investigate the adoption of Linked Data for commercial content providers from the requirements of online editors. The opportunities underlying Linked Data may fulfill their requirements as discussed in section 5.2. The Semtex application is aimed at supporting online editors during their editorial process and beyond. Results of a preliminary evaluation are presented in section 5.8 and showed that the SemTex application is already able to outperform other solutions. When implementing the first prototype, I confrontate a lot about challenges impairing a successful adoption of Linked Data for professional production environments, which are briefly outlined in section 5.9.

Chapter 6

Outlook

6.1 Summary on presented work

6.1.1 Chapter: Introduction

Introduction chapter outlined the motivation and circumstances considering the value of consuming and interlinking of data from heterogeneous Web data platforms in order to produce valuable and up-to-date online content that lead to the idea of this work. Definitions of problems that should be solved about the work of this thesis will be presented. The complexity of this work was summarized into five main objectives of matter that should be reached in the realm of proposed work. The chapter finally concludes with a brief outlook on the following work.

6.1.2 Chapter: Terminology extraction

This chapter gave us a short introduction of terminology extraction using two main approaches of NLP; linguistic and statistical. It also introduced SPPC and KEA, two mostly used tools that can be apply to extract german and english terms from free-text documents. The first part of this thesis has been focused on identification of related terms from the text which is implemented based on mentioned NLP methods and tools.

6.1.3 Chapter: ExistingWorks on Semantic Text Structuring, Analysis, and Interlinking

This chapter presents shortly the currently existing applications for semantic enrichment. It shows aspects like semantic knowledge representation of the content on the World Wide Web, semantic search, content analysis and visual representation. This applications which are already existent reflects to the achievement of this thesis's work.

6.1.4 Chapter: SemTex Application

In this chapter the details referring to the Semtex aplication including introduction, scenarios using Linked Data, the system architectural design, use cases, implementation, evaluation and challenges were introduced. This chapter explained much about professional production of valuable and up-to-date online content as outlined in section 5.1. The intention of this work was to to investigate the adoption of Linked Data for commercial content providers from the requirements of online editors. The opportunities underlying Linked Data may fulfill their requirements as discussed in section 5.2. The Semtex application is aimed at supporting online editors during their editorial process and beyond. Results of a preliminary evaluation are presented in section 5.8 and showed that the SemTex application is already able to outperform other solutions. When implementing the first prototype, I confrontate a lot about challenges impairing a successful adoption of Linked Data for professional production environments, which are briefly outlined in section 5.9.

6.1.5 Chapter: Outlook

Preview of presented chapters together with the outlook of possible future trends and the work that could follow this idea is introduced in this chapter, reflecting and encircling the achievements of current work, leading the scope to the closing chapter that follows.

6.2 Ideas for Future Work

This thesis already includes a preliminary evaluation of SemTex prototype as well as a comparison against other systems for semantic content enrichement. In a next step, the first prototype of SemTex will be rigorously evaluated with professional online editors as probands, deriving further requirements for technical improvement. The most important task is to easily integrate the next version SemTex prototype into the content managemement system of a major Austrian content provider. Drupal is one of content managemement system which is mostly RDF oriented. Annotation of text in Drupal is simple and straightforward and it do not require any knowledge about the details of RDF and description logics. The integration of prototype will enable us to perform other tests of the tentative hypothesis as shown in Figure 6.1 that embedding Linked Data positively affects the number of page visits and session length and thus has a positive effect on content providers' revenues.

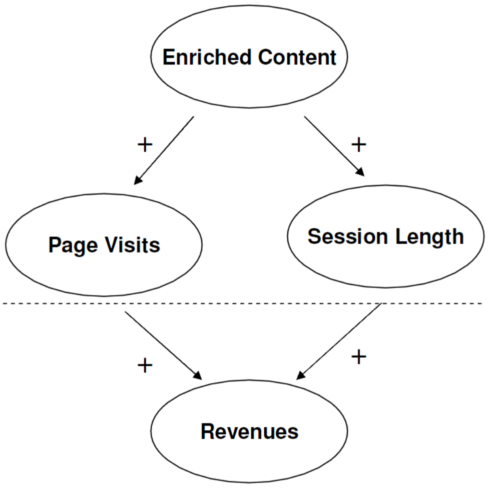


Figure 6.1: Tentative hypothesis of positive Linked Data

In current implementation are used only the two biggest datasets of Linked Open Data which include DBPedia and Geonames. DBpedia and Geonames contain sufficient semantically enriched information. Therefore, Semtex prototype consumes and interlinks its content only with these two datasets. Other datasets that are still on progress to fill their stores with RDF triples will be soon integrated in SemTex. The next main focus will be to discover further related data from media, publication and cross-domain datasets which can be challenging task. With semantic indexers such as Sindice or advertised SPARQL-end point is possible to get an idea what a dataset offers. But, in terms of scalability, conciseness, as well as convenience the above mentioned may not be the final solution.

An interesting task would be to extend the User Contributed Interlinking to multimedia assets. The current implemmentation of Flickr module is only about global metadata such as the creator, title or description. But Flickr offers more features such as for tagging and commenting. With these features we can focus on a fine-grained interlinking, for example, objects in a picture. In flickr it is possible to tag parts of a picture using so called ”notes”. The Flickr notes contain a string stating, e.g., ”X person is depicted in this picture” or ”X location is depicted in this picture”. Using these tags our Flickr module can extended in two steps. First, we can use flickr tags for matching people and locations which are already interlinked with Linked Open Data. Second, we retrive all pictures containing the desired person or location. CaMiCatzee from Hausenblas and Halb [2008] would be an example for the extension of Flickr module.

Bibliography

[2010]. *About Microformats*. `http://microformats.org/about`. (Cited on pages 20 and 21.)

Ben Adida and Mark Birbeck [2008]. *RDFa Primer*. `http://www.w3.org/TR/xhtml-rdfa-primer/`. (Cited on pages 19 and 86.)

A.J.Gerber, A. Barnard, and A.J.Van der Merwe [2007]. *Towards A Semantic Web Layered Architecture*. (Cited on page 10.)

K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao [2009]. *Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'*. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*. Madrid, Spain. (Cited on page 66.)

Sören Auer, Chris Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives [2008]. *DBpedia: A Nucleus for a Web of Open Data*. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, *Lecture Notes in Computer Science*, volume 4825, pages 722–735. Springer. (Cited on page 66.)

Ken Barker and Nadia Cornacchia [2000]. *Using Noun Phrase Heads to Extract Document Keyphrases*. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 40–52. AI '00, Springer-Verlag, London, UK. ISBN 3-540-67557-4. `http://dl.acm.org/citation.cfm?id=647461.726264`. (Cited on page 54.)

François Bellea, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette [2008]. *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems*. pages 41(5):706–16. (Cited on page 29.)

Tim Berners-Lee [2000]. *Semantic Web - XML*. `http://www.w3.org/2000/Talks/1206-xml2k-Tbl`. (Cited on page 10.)

Tim Berners-Lee [2003]. *The Semantic Web and Challenges*. `http://www.w3.org/2003/Talks/01-sweb-Tbl/overview.html`. (Cited on page 10.)

Tim Berners-Lee [2005]. *Web for real people*. `http://www.w3.org/2005/Talks/0511-keynote-tbl/`. (Cited on page 10.)

Tim Berners-Lee [2006]. *Artificial Intelligence and The Semantic Web*. `http://www.w3.org/2006`. (Cited on page 10.)

Tim Berners-Lee [2009]. *Weaving the Web*. HarperOne; 1st edition. (Cited on page 19.)

Tim Berners-Lee [September 1998]. *Semantic Web Road Map*. (Cited on page 10.)

Tim Berners-Lee, James Hendler, and Ora Lassila [2001]. *The Semantic Web*. Vol. 285, pages 28–37. (Cited on page 28.)

Chris Bizer and Richard Cyganiak [2009]. *D2R Server - Publishing Relational Databases on the Web as SPARQL-Endpoints*. `http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/`. (Cited on page 26.)

Chris Bizer, Tom Heath, and u.a [2007a]. *Interlinking Open Data on the Web*. www4.wiwiss.fu-berlin.de/bizer/pub/LinkingOpenData.pdf. `www4.wiwiss.fu-berlin.de/bizer/pub/LinkingOpenData.pdf`. Stand 12.5.2009. (Cited on page 66.)

Christian Bizer, Richard Cyganiak, and Tobias Gauss [2007b]. *The RDF Book Mashup: From Web APIs to a Web of Data*. In *Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW2007)*. (Cited on page 29.)

Christian Bizer, Tom Heath, and Tim Berners-Lee [2009]. *Linked Data - The Story So Far*. In *International Journal on Semantic Web and Information Systems, Special Issue on Linked Data*. `http://linkeddata.org/docs/ijswis-special-issue`. (Cited on page 28.)

Uldis Bojars, Alexandre Passant, and John G. Breslin [2008]. *Data Portability with SIOC and FOAF*. Dublin, Ireland. `http://2008.xtech.org/public/asset/attachment/378`. (Cited on page 23.)

Dan Brickley and R.V. Guha [2004]. *RDF Vocabulary Description Language 1.0: RDF Schema*. In *W3C Recommendation 10 February 2004*. `http://www.w3.org/TR/rdf-schema/`. (Cited on page 14.)

Stephane Corlosquet, Richard Cyganiak, Axel Polleres, and Stefan Decker []. *RDFa in Drupal: Bringing Cheese to the Web of Data*. (Cited on page 56.)

Ian Davis [July 2009]. *Linked data and the public domain*. `http://blogs.talis.com/nodalities/2009/07/linked-data-public-domain.php`. (Cited on page 92.)

Orri Erling [2008]. *Towards Web Scale RDF*. `http://www.openlinksw.com/uda/wiki/main//Main/VOSArticles/VOSArticleWebScaleRDF.pdf`. (Cited on page 24.)

D. Fensel [2000]. *The Semantic Web and Its Languages IEEE Intelligent Systems*. Vol.15 (6), pages 67–73. (Cited on page 10.)

D. Fensel [2002]. *Languages Standardization for the Semantic Web: The Long Way from OIL to OWL*. Vol. 2468, pages 215–227. (Cited on page 10.)

Roy T. Fielding and Richard N. Taylor [2002]. *Principled design of the modern Web architecture*. *ACM Trans. Internet Technol.*, 2(2), pages 115–150. ISSN 1533-5399. doi:http://doi.acm.org/10.1145/514183.514185. (Cited on page 64.)

Roy Thomas Fielding [2000]. *Architectural styles and the design of network-based software architectures*. ISBN 0-599-87118-0. Chair-Taylor, Richard N. (Cited on page 64.)

Jody Foo [2009]. *Term extraction using machine learning*. (Cited on page 6.)

W3C OWL Working Group [2009]. *OWL 2 Web Ontology Language*. `http://www.w3.org/TR/owl2-overview/`. (Cited on page 15.)

W. Halb, Y. Raimond, and M. Hausenblas [2008]. *Building Linked Data For Both Humans and Machines*. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*. Beijing, China. (Cited on page 51.)

Wolfgang Halb, Alexander Stocker, Harald Mayer, Helmut Muelner, and Ilir Ademi [2010]. *Towards a Commercial Adoption of Linked Open Data for Online Content Providers*. In *Proceedings of 6th International Conference on Semantic Systems (I-SEMANTICS)*. (Cited on pages v, 3, 44, 45 and 46.)

Oktie Hassanzadeh and Mariano Consens [2009]. *Linked Movie Data Base*. In *Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009)*. (Cited on pages 29 and 30.)

Michael Hausenblas [2009]. *Linked Data Applications*. (Cited on pages v, 48 and 49.)

Michael Hausenblas and Wolfgang Halb [2008]. *Interlinking Multimedia Data*. `http://triplify.org/Challenge/Nominations/files?get=hausenblas_camicatzee.pdf`. (Cited on page 97.)

Tom Heath and Enrico Motta [2008]. *Revyu: Linking reviews and ratings into the Web of Data*. In *Journal of Web Semantics*, pages 6(4):266–273. (Cited on page 29.)

Kenneth Holmqvist, Jana Holsanova, Maria Barthelson, and Daniel Lundqvist [2003]. *Reading or scanning? A study of newspaper and net paper reading*. (Cited on page 43.)

Ian Horrocks, Bijan Parsia, Peter Patel-Schneider, and James Hendler [September 2005]. *Semantic Web Architecture: Stack or Two Towers? In Third Workshop on Principles and Practice of Semantic Web Reasoning*. Dagstuhl,Germany. (Cited on pages v and 10.)

Ian Horrocks, F. Peter, and Patel-Schneider [2003]. *Three Theses of Representation in the Semantic Web*. ACM 1581136803, WWW2003, Budapest, Hungary. (Cited on page 10.)

Anja Jentzsch, Bo Andersson, Oktie Hassanzadeh, Susie Stephens, and Christian Bizer [2009]. *Enabling Tailored Therapeutics with Linked Data*. In *Proceedings of the 2nd Workshop on Linked Data onthe Web (LDOW2009)*. (Cited on page 29.)

Matthew Fisher John Hebeler, Ryan Blace, and Andrew Perez-Lopez [2009]. *Semantic Web Programming*. Wiley Publishing, Inc. (Cited on pages 9, 12, 16, 17, 23 and 28.)

Ali Khalili and Soren Auer [2011]. *The RDFa Content Editor From WYSIWYG to WYSIWYM*. `svn.aksw.org/papers/...RDFaEditor/public.pdf`. (Cited on page 56.)

Micheal Kifer [2008]. *Rule Interchange Format: The Framework in: Web Reasoning and Rule Systems*. `http://www.springerlink.com/content/e7v2802743688216/`. (Cited on page 16.)

Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee [2009]. *Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections*. In *Proceedings of the 6th European Semantic Web Conference (ESWC2009)*. (Cited on pages 29 and 46.)

Rob Larson and Evan Sandhaus [2009]. *Nyt to release thesaurus and enter linked data cloud*. `http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud`. (Cited on page 46.)

Atif Latif, Anwar Us Saeed, Patrick Hoefler, Alexander Stocker, and Claudia Wagner [2009]. *The Linked Data Value Chain: A Lightweight Model for Business Engineers*. pages 568–575. In Proceedings of I-SEMANTICS '09 International Conference on Semantic Systems. (Cited on pages v, 45 and 47.)

Julie B. Lovins [1968]. *Development of a stemming algorithm. Mechanical Translation and Computational Linguistics*, 11, pages 22–31. (Cited on page 8.)

Kelly Lyons, Corrie Playford, Paul R. Messinger, Run H. Niu, and Eleni Stroulia [2009]. *Business Models in Emerging Online Services*. In *Value Creation in E-Business Management. 15th Americas Conference on Information Systems, AMCIS 2009, SIGeBIZ track. Selected Papers, volume 36 of Lecture Notes in Business Information Processing*, pages 44–55. Springer Berlin Heidelberg, San Francisco, CA, USA. (Cited on pages 43 and 92.)

F. Manola and E. Miller [2007]. *RDF Primer*. http://www.w3.org/TR/rdf-primer/. (Cited on page 11.)

Andrew Matthews [2008]. *Understanding SPARQL - Create journaling micro-blogs with the semantic Web*. https://www6.software.ibm.com/developerworks/education/x-sparql/x-sparql-pdf.pdf. (Cited on page 17.)

Diana Maynard, Yaoyong Li, and Wim Peters [2008]. *NLP Techniques for Term Extraction and Ontology Population*. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 107–127. IOS Press, Amsterdam, The Netherlands, The Netherlands. ISBN 978-1-58603-818-2. (Cited on pages 5 and 69.)

Deborah L. McGuinness and Frank van Harmelen [2004]. *OWL Web Ontology Language*. http://www.w3.org/TR/owl-features/. (Cited on page 15.)

Alexander Osterwalder and Yves Pigneur [2002]. *An eBusiness Model Ontology for Modeling eBusiness*. In *In Proceedings of 15th Bled Electronic Commerce Conference. e-Reality: Constructing the e-Economy*. Bled, Slovenia 2002. (Cited on page 43.)

C. Paice and W.J. Black [2003]. *A three-pronged approach to the extraction of key terms and semantic roles*. (Cited on page 54.)

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto []. *Terminology extraction: an analysis of linguistic and statistical approaches*. (Cited on page 6.)

Jakub Piskorski and Günter Neumann [2000]. *An Intelligent Text Extraction and Navigation System*. In *RIAO*, pages 1015–1032. (Cited on pages v, 6 and 7.)

Eric Prud'hommeaux and Andy Seaborne [2008]. *SPARQL Query Language for RDF W3C Recommendation 15 January 2008*. http://www.w3.org/TR/rdf-sparql-query/. (Cited on pages 17 and 91.)

François Scharffe and Jerome Euzenat [2009]. *Alignments for data interlinking: Analysed systems*. http://melinda.inrialpes.fr/systems.html. (Cited on page 30.)

Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz [2007]. *A Practical OWL-DL Reasoner*. In *Journal of Web Semantics*. http://pellet.owldl.com/papers/sirin07pellet.pdf. (Cited on page 27.)

Yasin Uzun []. *Keyword Extraction Using Naive Bayes*. (Cited on page 7.)

Julius Volza, Christian Bizer, Martin Gaedke, and Georgi Kobilarov [2009]. *Silk - A Link Discovery Framework for the Web of Data*. In *In: WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*. (Cited on page 30.)

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning [1999]. *KEA: practical automatic keyphrase extraction*. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255. ACM, New York, NY, USA. ISBN 1-58113-145-3. doi:http://doi.acm.org/10.1145/313238.313437. (Cited on pages v, 7 and 8.)

Fabio Massimo Zanzotto [2002]. *L'estrazione della terminologia come strumento per la modellazione di domini conoscitivi*. (Cited on page 5.)