



**Project Acronym: Europeana v2**  
**Grant Agreement number: 270902**  
**Project Title: Europeana Version 2**

## **D7.4: Market study on technical options for semantic feature extraction**

<b>Revision</b>	Final
<b>Date of submission</b>	27.04.2012
<b>Author(s)</b>	Marlies Olensky, Humboldt-Universität zu Berlin
<b>Dissemination Level</b>	[Public]

Project co-funded by the European Commission within the ICT Policy Support Programme

**REVISION HISTORY AND STATEMENT OF ORIGINALITY****Revision History**

<b>Revision No.</b>	<b>Date</b>	<b>Author</b>	<b>Organisation</b>	<b>Description</b>
Draft 1	16.02.2012	Marlies Olensky	Humboldt-Universität zu Berlin	First draft version
Draft 2	19.03.2012	Marlies Olensky	Humboldt-Universität zu Berlin	Incorporating feedback from V. Charles
Draft 3	16.04.2012	Marlies Olensky	Humboldt-Universität zu Berlin	Finalizing the “text tools part”, adding the “multimedia tools part” and incorporating feedback from M. Brinkerink
Final	27.04.2012	Marlies Olensky	Humboldt-Universität zu Berlin	Incorporating feedback from S. Gradmann, V. Charles and A. Isaac
Final	30.04.2012	Marlies Olensky	Humboldt-Universität zu Berlin	Incorporating feedback from S. Gradmann

**Statement of originality:**

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

## Contents

Contents.....	3
Scope of this study .....	4
What is semantic extraction?.....	5
Tools for semantic extraction - text.....	6
AlchemyAPI.....	6
Apache Stanbol .....	8
CiceroLite.....	9
DBpedia Spotlight .....	10
Evri.....	11
Luxid® for Content Enrichment .....	12
Ontonaut.....	13
OpenCalais .....	13
Open Sahara .....	15
PoolParty .....	15
Sophia Semantic Engine.....	16
Sw2sws.....	16
Unlock Text.....	17
Wikimeta .....	17
Zemanta.....	18
Feature Matrix .....	20
Tools for semantic extraction – multimedia.....	22
Annnotation and SugarTube.....	22
Automatic semantic video indexing of Turkish news videos .....	23
Contentus.....	24
Impala.....	25
KAT (K-Space Annotation Tool).....	26
Semantic Video Annotation Suite .....	26
References .....	28
Appendix .....	30
Demo texts.....	30
Resources .....	32
Other exemplary state-of-the-art tools .....	33

## Scope of this study

This market study comprises an overview of available tools that can perform semantic extraction. In this way it continues and complements work from Europeana v1.0 and EuropeanaConnect. The requirement was to investigate effective and efficient tools for the extraction of concepts and named entities from digital text, audio and video resources. Consequently the tools should enable the automatic semantic contextualization for object metadata lacking contextualization links. We based our study on the TELplus deliverable on “State of the art of semantic and multilingual engines or tools for digital libraries” (Freire, Mane & Petz, 2008) and aimed at complementing it with additional tools that perform semantic extraction. That is why, for example the framework GATE<sup>1</sup> is not included in this study. The selection criteria for our study are specified in the following paragraphs.

The selection criteria for the tools extracting semantics from text were the following: They should have at least a web-service available where the document to be contextualized can be uploaded in order to receive an enriched document as output. The tools should perform named entity recognition and be based on one or more lexical or structured sources. They could be either commercial, free with limits on the call quota or open source. Wherever possible, we tested the demo with a text about the Berlin Cathedral from Wikipedia. The demo texts can be found in the Appendix with a paragraph showing what entities could be recognized by the tools. The tests should only be considered as an example of what functionality the tools actually can provide and give a first impression of how complete the information extraction is. Yet, when considering a partner for Europeana more in-depth testing will be necessary, similar to the experiments made by Rizzo & Troncy (2011) and Rizzo, Troncy, Hellmann & Bruemmer (2012).

The tools that are able to extract entities and concepts from audio, video and images are still in an experimental state and mostly research in progress. The tools often do not fully exploit the extracted information by linking them to resources on the Semantic Web. Others only assist the user with annotating the specific media types and do not even perform semi-automatic semantic extraction. We therefore chose tools that either annotated multimedia content with Linked Data resources or provide at least semi-automatic semantic indexing or in the best case both. We did not consider tools assisting the user to manually annotate multimedia content with ontologies. Not in the scope of this study are tools that provide multimedia retrieval solely based on similarity and do not extract entities or concepts.

Not in the scope of this study is the tool Collexis that was taken over by Elsevier because it is now incorporated in SciVal: Elsevier Fingerprint Engine™, an expertise profiling and research networking tool (SciVal, 2012). This market study does not claim to be complete. Therefore, links to other exemplary state-of-the-art tools or frameworks can be found in the Appendix. Additionally, resources used to find suitable tools for the market study are listed in the Appendix.

Tools that perform semantic extraction on text are described in one chapter; those working on multimedia are summarized in a separate one.

---

<sup>1</sup> <http://gate.ac.uk/>

## **What is semantic extraction?**

Semantic extraction refers to extracting the meaning of a document, preferably in a (semi-) automated way. Part of semantic extraction are the following processes: recognizing and extracting named entities and keywords, analyzing the sentiment of a document, extracting facts and relation between those facts and named entities, categorizing documents, recognizing and extracting concepts and finally adding them as metadata or annotations.

Extracting semantics from images, audio and video objects is a more difficult procedure. The main challenges for audio files are the automatic recognition of speech, speakers, background noise and musical analysis. The problems for images and videos are the semantic and the sensory gap (Worring, 2008). The semantic gap refers to going beyond the descriptive attributes of an image and identifying a specific person as opposed to just identifying for example a female person with blonde hair. The sensory gap refers to distinguishing between similar objects and identifying the same ones. Technologies employed are automatic speech recognition, face detection, exploiting existing metadata, video text analysis, image and video segmentation and event extraction.

## Tools for semantic extraction - text

This chapter lists all tools in alphabetical order that perform semantic extraction on text matching our requirements defined in the chapter *Scope of this study*.

### AlchemyAPI

AlchemyAPI<sup>2</sup> can perform named entity extraction, keyword extraction, sentiment analysis, fact and relation extraction, document categorization, concept tagging, language detection, and structured content scraping. It employs the methods of deep linguistic parsing, statistical natural language processing, and machine learning. It can extract information about people, places, companies, topics and languages and this information as semantic metadata to the content and states to use a sophisticated 'entity disambiguation' mechanism. AlchemyAPI's entity extraction works on a specific set of defined entities (a few hundred, a complete list can be found online<sup>3</sup>). Named entity recognition works for the following languages: English, French, German, Italian, Portuguese, Russian, Spanish, and Swedish. Yet, the language detection function works for more than 95 languages. (AlchemyAPI, 2012)

AlchemyAPI can integrate content with a variety of resources from the Linked Data cloud. The following Linked Data resources are currently leveraged by AlchemyAPI: Freebase, US Census, GeoNames, UMBEL, OpenCyc, YAGO, MusicBrainz, CIA Factbook, CrunchBase. AlchemyAPI provides also quotations extraction, as well as the ability to extract entity-level sentiment, i.e. identify positive and negative statements (on document-level and keyword-level). Extracted metadata can be returned in different formats, including XML, JSON, RDF, and Microformats. Additionally AlchemyAPI performs identification of Subject-Action-Object relations within a HTML-document, text, or web-based content. (AlchemyAPI, 2012)

It works as an API, where 1,000 API calls a day are free of charge. If they are contacted, approved academic users can receive up to 30,000 API calls a day. Higher limits are obtainable for educational institutions and non-profit groups.

We tried the demo with the English demo text (all demo texts can be found in the Appendix) about the Berlin Cathedral. It did not recognize all of the possible entities. Yet, the ones that were identified were correctly classified. Figure 1 shows the results. The tool also recognized 7 relations that are displayed in Figure 2. The demo worked also for the German, French and Spanish texts, though it recognized fewer entities than in the English version. Additionally, no relations were identified in those other languages.

---

<sup>2</sup> <http://www.alchemyapi.com/>

<sup>3</sup> <http://www.alchemyapi.com/api/entity/types.html>

D7.4: Market study on technical options for semantic feature extraction

**AlchemyAPI**  
Transforming Text Into Knowledge

Home Documentation Tools Products Blog Company

**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!

St. Hedwig's Cathedral **serves** as seat of Berlin's Roman Catholic metropolitan bishop

[Click here to go back to the text entry page](#)

**Berlin Cathedral** (German: [Berliner Dom](#)) is the colloquial name for the Evangelical (i.e. Protestant) Oberpfarr- und Domkirche (English analogously: [Supreme Parish](#) and [Collegiate Church](#), literally [Supreme Parish](#) and [Cathedral Church](#)) in [Berlin, Germany](#). It is the parish church of the Evangelical congregation Gemeinde der Oberpfarr- und Domkirche zu [Berlin](#), a member of the umbrella [organisation Evangelical Church](#) of Berlin-Brandenburg-Silesian Upper Lusatia. Its present building is located on [Museum Island](#) in the Mitte borough.

The [Berlin Cathedral](#) had never been [a cathedral](#) in the actual sense of that term since it has never been the seat of a bishop. The bishop of the [Evangelical Church in Berlin-Brandenburg](#) (under this name 1945–2003) is based in [St. Mary's Church, Berlin](#), and Kaiser [Wilhelm Memorial Church](#). [St. Hedwig's Cathedral](#) serves as seat of [Berlin's](#) Roman Catholic metropolitan bishop.

**Language**  
english

**Facility (4)**  
[Berlin Cathedral](#)  
[St. Hedwig's Cathedral](#)  
[Wilhelm Memorial Church](#)  
[St. Mary's Church](#)

**City (2)**  
[Berlin](#)  
[Berlin-Brandenburg](#)

**Organization (4)**  
[organisation Evangelical](#)

Figure 1. AlchemyAPI Demo

**Relations (7) [hide](#)**

Subject	Action	Object
<a href="#">the colloquial name for the Evangelical (i.e</a>	is	<a href="#">Berlin Cathedral (German</a>
<a href="#">It</a>	is	<a href="#">the parish church of the Evangelical congregation Gemeinde der Oberpfarr- und Domkirche zu Berlin</a>
<a href="#">Its present building</a>	is located	
<a href="#">The Berlin Cathedral</a>	had never been	<a href="#">a cathedral</a>
<a href="#">it</a>	has never been	<a href="#">the seat of a bishop</a>
<a href="#">The bishop of the umbrella organisation Evangelical Church</a>	is based	
<a href="#">St. Hedwig's Cathedral</a>	<b>serves</b>	<a href="#">as seat of Berlin's Roman Catholic metropolitan bishop</a>

Figure 2. AlchemyAPI Demo, relations

## Apache Stanbol

Apache Stanbol<sup>4</sup> has incorporated the FISE Semantic Engine<sup>5</sup> into its project. The tool is an open source set of components semantically analyzing and enriching content. It can be installed locally or used via the API. The employed technologies are natural language processing, metadata extraction and linking named entities to public or private entity repositories. Also the tool then applies rules and reasoning to add additional knowledge and links. (Apache Stanbol, 2012) In addition, it is possible to use other components like the content hub (allows storing results on the server) or the fact store (allows storing facts, i.e. semantic relations between entities). A full list of components and their functions can be found on the website<sup>6</sup>.

We tested the demo<sup>7</sup> with the English demo text. As output format we could choose between: JSON-LD, RDF/XML, RDF/JSON, Turtle and N-Triples. The actual extraction and contextualization was able to extract entities matching organizations and places, but did only find a few, including one that did not make sense at all (cf. Figure 3, St. Louis, Missouri). Additionally it showed the places on a map. Figure 3 shows the found matches; Figure 4 shows the RDF/XML output.

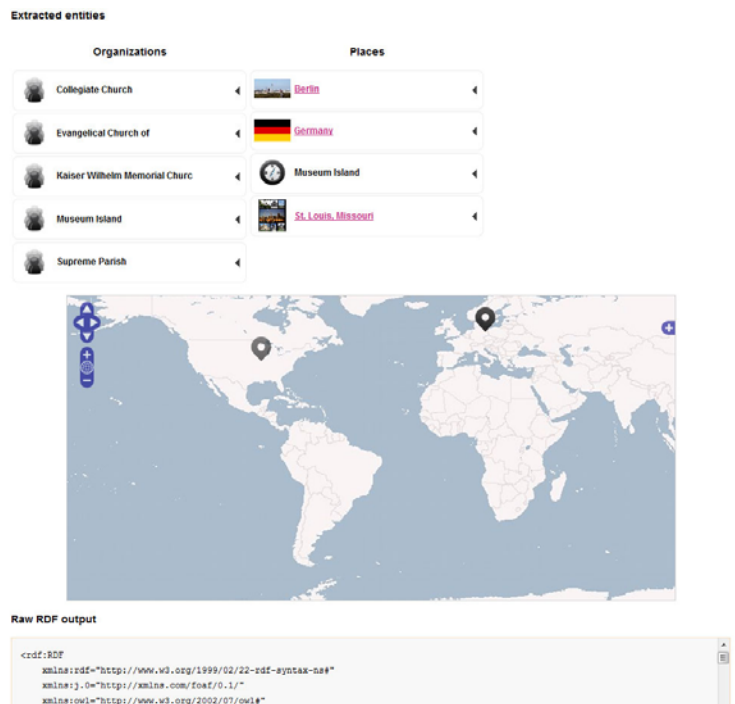


Figure 3. Apache Stanbol Demo, extracted entities

<sup>4</sup> <http://incubator.apache.org/stanbol/index.html>

<sup>5</sup> <http://wiki.iks-project.eu/index.php/FISE>

<sup>6</sup> <http://incubator.apache.org/stanbol/docs/trunk/components.html>

<sup>7</sup> <http://stanbol.demo.nuxeo.com/engines>



Raw RDF output

```

<j.7:start rdf:datatype="http://www.w3.org/2001/XMLSchema#int">350</j.7:start>
<j.7:confidence rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.8871081533117463</j.7:confidence>
<j.1:type rdf:resource="http://dbpedia.org/ontology/Place"/>
<j.7:selection-context rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Protestant Oberpfarr- und Domkirche (English analogously: Supreme Parish and Collegiate Church, literally Supreme Parish and Cathedral Church) in Berlin, Germany. It is the parish church of the Evangelical congregation Gemeinde der Oberpfarr- und Domkirche zu Berlin, a member of the umbrella organisation Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia. Its present building is located on Museum Island in the Mitte borough.

The Berlin Cathedral had never been a cathedral in the actual sense of that term since it has never been the seat of a bishop.
</j.7:selection-context>
<j.7:selected-text rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Berlin</j.7:selected-text>
<rdf:type rdf:resource="http://fise.lks-project.eu/ontology/TextAnnotation"/>
<j.1:creator rdf:datatype="http://www.w3.org
/2001/XMLSchema#string">org.apache.stanbol.enhancer.engines.opennlp.impl.NEREngineCore</j.1:creator>
<j.1:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2012-04-10T11:08:11.071Z</j.1:created>
<j.7:extracted-from rdf:resource="urn:content-item-sha1-1f7cf0ca2a98e440f23e42edaaf18c76e06754e1"/>
<rdf:type rdf:resource="http://fise.lks-project.eu/ontology/Enhancement"/>
</rdf:Description>
<rdf:Description rdf:about="urn:enhancement-a3b7843b-9957-9078-024c-1138fec2cb2">
<j.7:selection-context rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Berlin Cathedral (German: Berliner Dom) is the colloquial name for the Evangelical (i.e. Protestant) Oberpfarr- und Domkirche (English analogously: Supreme Parish and Collegiate Church, literally Supreme Parish and Cathedral Church) in Berlin, Germany. It is the parish church of the Evangelical congregation Gemeinde der Oberpfarr- und Domkirche zu Berlin, a member of the umbrella organisation Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia.</j.7:selection-context>
<j.7:selected-text rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Germany</j.7:selected-text>
<rdf:type rdf:resource="http://fise.lks-project.eu/ontology/TextAnnotation"/>
<j.1:creator rdf:datatype="http://www.w3.org
/2001/XMLSchema#string">org.apache.stanbol.enhancer.engines.opennlp.impl.NEREngineCore</j.1:creator>
<j.1:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2012-04-10T11:08:11.071Z</j.1:created>
<j.7:extracted-from rdf:resource="urn:content-item-sha1-1f7cf0ca2a98e440f23e42edaaf18c76e06754e1"/>

```

Figure 4. Apache Stanbol Demo, RDF output

## CiceroLite

Language Computer's CiceroLite<sup>8</sup> used to be known as Extractiv's Semantic on Demand (a NLP-technology to attach semantic metadata, like entities and relations, to the content submitted). On Extractiv's website it says that the service is no longer available, however when directly going to the Demo page<sup>9</sup> it still seems to work. Yet, in this study we concentrate on CiceroLite that apparently has incorporated Extractiv's technology for named entity recognition.

CiceroLite states to work for English, Modern Standard Arabic, Mandarin Chinese, Japanese, Spanish, German and Dutch. Due to its machine learning framework, it is also extensible to new languages when sources of training data are available. It indicates to be able to identify more than 150 types of named entities in English with a precision of more than 90%. These entity types can be varied according to customer needs and expanded to up to 250 domain-specific entities. (Language Computer, 2012)

We tested the demo<sup>10</sup> with the German text from Wikipedia about the Berlin Cathedral, but about a third of the entities were assigned to a wrong entity type. Some of the recognized entities included links to dbpedia. The English version showed better result and most of the entities were classified correctly. Yet, some entities were not identified correctly. For example St. Mary's Church was not identified as one but split into the location "St. Mary" and the organization "church". Also, the word cathedral was identified as a commercial organization (cf. Figure 5).

CiceroLite needs to be licensed. Costs need to be inquired and would be calculated according to customers' needs.

<sup>8</sup> <http://www.languagecomputer.com/products/text-annotation/cicerolite.html>

<sup>9</sup> <http://www.extractiv.com/demo.html>

<sup>10</sup> <http://demo.languagecomputer.com/cicerolite>

The screenshot displays the CiceroServer: English interface. At the top, there are navigation links for 'Language Computer', 'Contact Us', and 'Feedback'. The main header features the 'CiceroServer: English' logo and the 'LANGUAGE COMPUTER' logo with a stylized 'i' icon. The interface is divided into several sections:

- ENTITIES:** A list of entity types with counts: LOCATION (15), ORGANIZATION (7), OTHER (9), and DATE-TIME (2). There are 'Expand All' and 'Collapse All' options.
- RELATIONS:** A list of relation types with counts: GENERIC (7). There are 'Expand All' and 'Collapse All' options.
- DOCUMENT FILE:** A text area containing the document content. The text is annotated with colored boxes around specific words and phrases, such as 'Berliner Dom', 'Supreme Parish and Collegiate Church', 'Evangelical Church', and 'Kaiser Wilhelm Memorial Church'.
- ENTITY TYPE KEY:** A legend showing color-coded boxes for PERSON (blue), LOCATION (green), ORGANIZATION (red), CONTACT\_INFO (orange), OTHER (purple), DATE-TIME (pink), and NUMERIC (dark green).
- DETAILS:** A section titled 'Document MetaData' showing:
  - topics: autos\_&\_vehicles: 72%
  - doctype: HTML
  - language: ENGLISH
  - raw\_size: 909

Additional text on the right side of the interface includes: 'Mouse-over entities and relations to see more detailed type information about them.' and 'Additional information about these types can be found in our online documentation (entities, relations)'.

Figure 5. CiceroServer Demo, English

## DBpedia Spotlight

DBpedia Spotlight<sup>11</sup> is a tool for annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. It works on text only and offers three functions: annotate, disambiguate and candidates (find candidate DBpedia resources). (Mendes, Jakob, Garcia-Silva et al., 2011)

DBpedia Spotlight can either be accessed through a web application (which is a demo client) where you can enter text and the tool will create an HTML-version of the text including the DBpedia annotations. Or you can use the Scala / Java API, REST Web Service to get the functionality of annotating and/or disambiguating entities in text. There are also two other APIs available: Annotation Java / Scala API, exposing the underlying logic that performs the annotation/disambiguation and Indexing Java / Scala API, executing the data processing necessary to enable the annotation/disambiguation algorithms used. More technical information can be found in the User Manual<sup>12</sup> or the Technical Documentation<sup>13</sup>. A local installation on one's own web server would also be possible. The necessary downloads can be found on DBpedia Spotlight's main page. The program can be used under the terms of the Apache License, 2.0. Part of the code uses LingPipe under the Royalty Free License. Therefore, this license also applies to the output of the currently deployed web service. (Mendes, Jakob, Daiber et al., 2011)

We tested the DBpedia Spotlight Demo<sup>14</sup> with the short English text about the Berlin Cathedral, as the demo only works in English. We set the parameters like described in the user's manual: confidence: 0,4; contextual score: 0,0; prominence (support): 20; no common words; document-centric and show n-best candidates. The result is shown in Figure 6.

<sup>11</sup> <http://dbpedia.org/spotlight>

<sup>12</sup> <http://wiki.dbpedia.org/spotlight/usersmanual?v=i0m>

<sup>13</sup> <http://wiki.dbpedia.org/spotlight/technicaldocumentation?v=3qy>

<sup>14</sup> <http://spotlight.dbpedia.org/demo/index.html>

Interestingly the tool found more concepts than those linked in Wikipedia, yet not all of them are identified correctly. For example in Wikipedia the Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia is recognized as one concept, in the demo it divides the concept into single parts. When changing the support parameter fewer entities are detected, yet again, the precision of Wikipedia is not reached.



Figure 6. DBpedia Spotlight Demo

Currently they only offer a web service for DBpedia Spotlight in English. However, since it's based on Wikipedia, one could use the DBpedia Spotlight software to build a service for any language that is in Wikipedia. There are some minor changes needed for the most basic features of the tool, and for using more NLP-intensive features, it needs a few more changes. They plan to do that at some point this year (2012), but it depends on project funding. (Mendes, 2012)

DBpedia Spotlight has also planned and prototyped a function that uses graph-structured metadata for disambiguation, but it is not yet finished. They would like to collaborate on that if Europeana has a real-world use case. (Mendes, 2012)

## Evri

Looking at Evri's main website<sup>15</sup> it looks as if Evri offers only apps for smartphones. Yet, the underlying technologies for these apps are text analysis as well as information and knowledge extraction. Figure 7 shows Evri's basic structure. The API can extract named entities, categorize them, identify relationships between entities and add recommendations for related content. Yet, the API is only available to developers contributing as testers and evaluators for free. Other application areas need to be negotiated with Evri. (Evri Developer Center, 2012)

<sup>15</sup> <http://www.evri.com/>

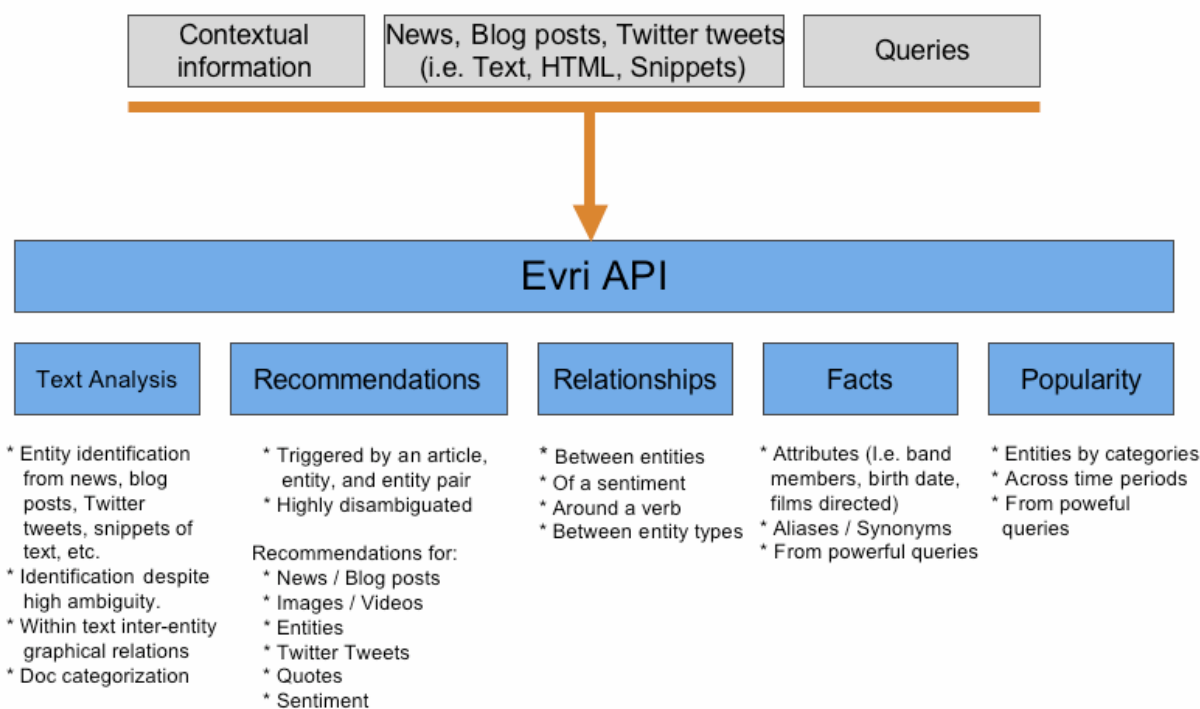


Figure 7. Evri's basic structure

Evri does not state what kind of technologies (like natural language processing) they employ. It also does not mention any multilingual capabilities. That is why we assume Evri works for the English language only. Unfortunately no demo was available online.

### Lucid® for Content Enrichment

Lucid®<sup>16</sup> is text mining software and can perform content tagging, linking, conceptual extraction, automatic classification and annotation. The Lucid® Annotation factory can apply taxonomies, thesauri and lexicons and the Knowledge Factory actually annotates content. Supported languages are English, French, German, Greek, Hungarian, Italian, Czech, Dutch, Polish, Portuguese, Russian and Spanish. Additional language packs are available for: Danish, Swedish, Finnish, Norwegian, Chinese, Korean, Japanese and Arabic. The product has to be licensed. (TEMIS, 2012)

Unfortunately no demo version was available online. The only tool to be tested is the Lucid Toolbar, which is a browser plug-in for Internet Explorer or Firefox which can only be installed under Windows. Yet, unfortunately it does not work under Windows Vista, which is the reason why we were not able to test the toolbar. To our best knowledge TEMIS is currently in negotiations with Springer.

In 2010/2011 we have already been in contact with TEMIS Germany and have started cooperating with them. We tried out their RTF tool on a small sample of data. This RTF tool basically annotates free-text with provided controlled vocabularies and provides a confidence-score for the annotations. It would be therefore worth investigating, if this cooperation could be continued and the tool be tested on a larger set of Europeana data and appropriate controlled vocabularies.

<sup>16</sup> <http://www.temis.com/index.php?id=30&selt=15>

## Ontonaut

Ontonaut<sup>17</sup> is a toolkit to be employed for websites, blogs, etc. in order to semantically enrich content. It was developed by two companies (Top 21<sup>18</sup> from Germany, Attensity<sup>19</sup> from the US but with German location) within the Theseus research program<sup>20</sup> and is currently in the public Beta phase. One needs to register for an API key and then can use the tool for up to 1,000 calls per day. Other usage needs to be negotiated with the companies. (Ontonaut, 2012)

Ontonaut provides two different services, called *Ontonaut extract* and *Ontonaut enrich*. Like the names say the *extract* service performs named entity extraction (no further specification of technologies is mentioned) and the *enrich* service provides background information based on a search in Freebase and DBpedia. It states it can automatically recognize the language of a text. Yet, currently Ontonaut only works for English and German texts. Unfortunately no demo was available online.

## OpenCalais

OpenCalais<sup>21</sup> is owned by ThomsonReuters; it provides a web service that automatically extracts semantic metadata from the content submitted. It uses natural language processing (NLP), machine learning and other methods. It is able to perform named entity recognition and also adds facts and events to them. OpenCalais creates metadata in RDF format that can then be used in other applications. OpenCalais mainly works on English text, but a limited number of entities can be extracted in the French and Spanish modules. OpenCalais can also be used as an API. Figure 8 shows OpenCalais' functionalities. (Open Calais, 2012)

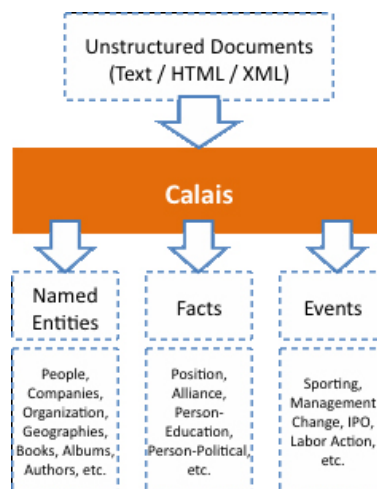


Figure 8. OpenCalais functionalities

The web service is free for commercial and non-commercial use but limited up to 50,000 transactions per day per user at a maximum rate of four transactions per second. For academic or other similar users they evaluate raising these limits on a case-by-case basis. OpenCalais is based on their own ontology. In each language the tool has a different range

<sup>17</sup> <http://ontonaut.net/>

<sup>18</sup> <http://www.top21.de/>

<sup>19</sup> <http://www.attensity.com/>

<sup>20</sup> <http://www.theseus-programm.de/en/index.php>

<sup>21</sup> <http://www.opencalais.com/>

D7.4: Market study on technical options for semantic feature extraction

of metadata that can be identified. The most comprehensive set is available for English. The specific features are listed in Table 1. (Open Calais, 2012)

Table 1. Features of API Metadata

Features	English	French	Spanish
Entities (e.g. Anniversary, City, Company, etc.)	X	X	X
Events and Facts (e.g. Acquisition, Alliance, AnalystEarningsEstimate, etc.)	X		
Generic Relation Extraction	X		
Entity Relevance Score	X	X	X
Social Tags (no real semantic extraction, rather common sense association)	X		
Document Categorization	X		
Entity Disambiguation	X		

We tested OpenCalais with the English demo text about the Berlin Cathedral from Wikipedia. The results are shown in Figure 9. The tool recognized most entities and also classified them correctly. When entering the text in German, the tool recognizes that the language is not supported and gives an error message. Trying the French text the tool still works fine regarding the classification, although it finds fewer entities, the Spanish text provides the least satisfying results (with even fewer entities and some entities classified incorrectly).

The screenshot displays the OpenCalais interface. At the top left is the logo 'CALAIS Powered by Thomson Reuters'. Below it are buttons for 'Show RDF' and 'Entry Page'. The main content area is divided into several sections:

- Topics:** Religion Belief (98%), Politics (97%).
- Social Tags:** A list of tags with star ratings, including Politics, Religion Belief, Berlin, Protestant churches, States of Germany, Prussian Union, Saint Thomas Church, St. Hedwig's Cathedral, Mitte, and Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia.
- Entities:** A list of entity types with checkboxes, including City, Country, Facility, Natural Feature, Organization, Position, Province Or State, and Technology.
- Events & Facts:** A section with a checkbox for 'Generic Relations'.

The main text area contains a snippet from Wikipedia about the Berlin Cathedral, with several terms highlighted in yellow to indicate entity recognition. The highlighted text includes: 'Berlin Cathedral', 'Berliner Dom', 'Supreme Parish and Collegiate Church', 'parish church', 'umbrella organisation Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia', 'Museum Island', 'Mitte borough', 'The Berlin Cathedral had never been a cathedral', 'bishop', 'Evangelical Church in Berlin-Brandenburg', 'St. Mary's Church', 'Berlin', 'Kaiser Wilhelm Memorial Church', 'St. Hedwig's Cathedral', and 'metropolitan bishop'.

Figure 9. OpenCalais Demo

## Open Sahara

Open Sahara<sup>22</sup> is a free web service (also available as API) for semantically enriching documents, websites and blogs. The main target groups are bloggers, journalists and software developers. The technologies employed are smart language processing and statistical algorithms. Open Sahara can extract entities, facts, and events. (Open Sahara, 2012)

Unfortunately it does not say on what content it bases the semantic extraction. Also, the supported languages are not specified on their website. Yet, we assumed the supported languages are Dutch and English, as the University of Amsterdam is involved and the website is available in these two languages. No demo was available online.

## PoolParty

PoolParty<sup>23</sup> is a product family developed by Semantic Web Company. The three components of PoolParty are the Thesaurus Manager, the Extractor and the Semantic Search. The Thesaurus Manager maintains all kinds of controlled vocabularies in different formats (SKOS, RDF, SPARQL) and provides functionality for publishing the vocabularies as linked open data. The Extractor provides text mining algorithms based on semantic knowledge models and can extract meaningful phrases, named entities, categories or other metadata. (Semantic Web Company, 2012)

Figure 10. PoolParty Extractor Demo

<sup>22</sup> <http://opensahara.com/>

<sup>23</sup> <http://poolparty.biz/>

We tested the demo<sup>24</sup> with the demo text and it gave reasonable results for English and German (cf. Figure 10). The entities are not marked in the text, but given in the recommended tags at the bottom of the page. Not all of the recommended tags would be suitable to add to the text, but the recommended images match with the text. The recommended documents at the first ranks are useful links to related content.

PoolParty is quite flexible with languages. The Extractor is basically language-independent and can be fed with thesauri (in SKOS format) in any language. To further improve the Extractor it can also integrate stop word lists from the specific language. (Blumauer, 2012) The PoolParty Extractor can be installed behind one's own firewall or used as a web service. The price is Euro 11.000,- per server instance and there are no CPU limits. They include also additional test/preprod-license free of charge and an optional license maintenance fee can be obtained for 30% of the server license per year. (Semantic Web Company, 2012)

### Sophia Semantic Engine

The Sophia Semantic Engine<sup>25</sup> is commercial software developed by the Italian company CELI. It works with natural language processing and is able to identify entities like events, persons' names, places and brands. In addition, it can classify and annotate documents and also extract domain-specific technical terms. The software is based on proprietary morpho-syntactic analysis modules and on a broad semantic lexicon. CELI does not specify the supported languages, but we assume that it works for languages where language resources are available (as CELI has previously worked on multilinguality for Europeana). Also we assumed the software works only as local installation and CELI does not offer an API or web service. Unfortunately no demo was available online.

### Sw2sws

Sw2sws<sup>26</sup> (Sitio Web en un Sitio Web Semántico) is a tool that can create simple annotations from text in Spanish. The tool is an experimental tool that can be used under the EUPL V.1.1 license<sup>27</sup>. You can download the tool from the sw2sws website. (Criado-Fernandez, 2010)

The tool is based on five ontologies:

- <http://www.daml.org/2001/01/gedcom/gedcom>
- <http://www.w3.org/2001/sw/WebOnt/guide-src/food>
- <http://www.w3.org/2001/sw/WebOnt/guide-src/wine>
- <http://www.co-ode.org/ontologies/pizza/2005/05/16/pizza.owl>
- [http://www.criado.info/owl/vertebrados\\_es.owl](http://www.criado.info/owl/vertebrados_es.owl)

It works only on Spanish web content and is still in a very experimental stage.

---

<sup>24</sup> <http://poolparty.biz/demozone/>

<sup>25</sup> <http://www.celi.it/en/sophia-semantic-engine.shtml>

<sup>26</sup> <http://sw2sws.sourceforge.net/>

<sup>27</sup> <http://joinup.ec.europa.eu/software/page/eupl>



## Unlock Text

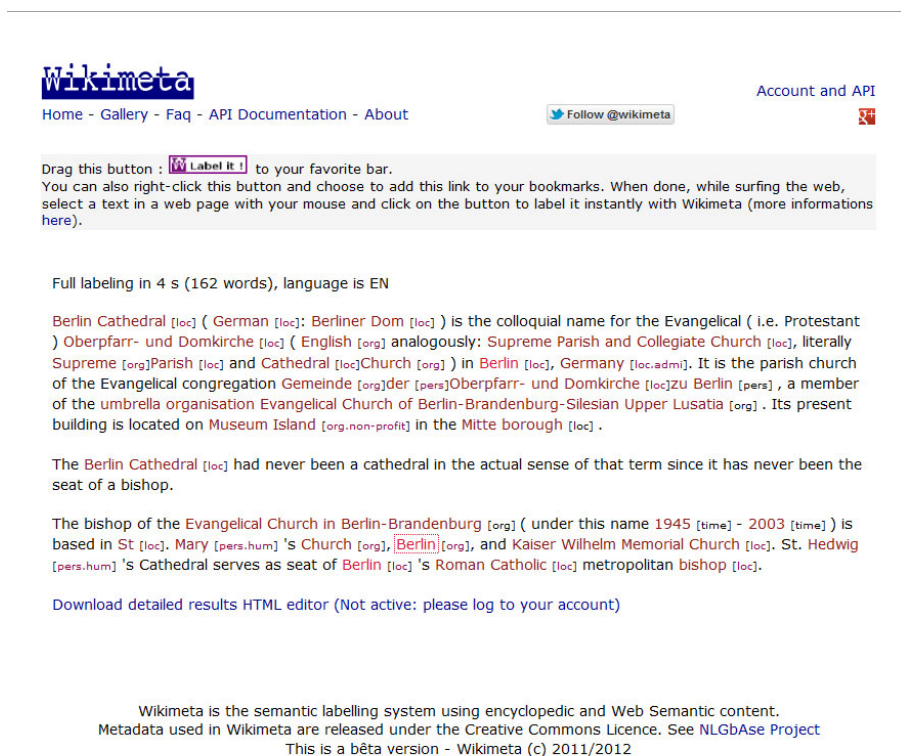
Unlock Text<sup>28</sup> was developed by EDINA which is a JISC National Data Centre based at the University of Edinburgh. It is a geo-parsing tool that can extract geo-related information from documents. Unlock Text is a RESTful API that uses another API (Unlock Places API) to generate a geo-referenced set of results. The text you would like to have enriched needs to be available as an html or txt document on the web. Unlock Text is based on GeoNames<sup>29</sup> and other web content that is not further specified. (Unlock, 2012)

Language wise the tool works only on English texts. Yet, the matching service from Unlock Places in principal works on all name variations and translations that are found in GeoNames. (Walsh, 2012) Unfortunately no demo was available online.

## Wikimeta

Wikimeta<sup>30</sup> is an academic research project from the École Polytechnique de Montréal. Wikimeta is a tool for automatically annotating text and linking it to the Linked Open Data cloud. It uses DBpedia, Geonames, CIA World Factbook or the web if no other matching resource can be found. It also performs named entity recognition prior to searching for links in the LOD cloud. It is still in beta version. Wikimeta claims to be fully multilingual but actually works only on English and French text. Yet, they do use language-specific disambiguation models. They plan on including more languages. In principle Wikimeta has a day limit on the API. The quota is 100 calls and 1 megabyte of data a day but if a larger amount is needed, Wikimeta can be contacted for further negotiation. (Wikimeta, 2012)

---



The screenshot shows the Wikimeta website interface. At the top, there is a navigation bar with links for Home, Gallery, Faq, API Documentation, and About. A 'Follow @wikimeta' button is also present. The main content area features a 'Label It!' button and instructions on how to use it. Below this, a text snippet is shown with semantic labels applied to various words and phrases, such as 'Berlin Cathedral', 'German', 'Berliner Dom', 'Evangelical', 'Protestant', 'Oberpfarr- und Domkirche', 'English', 'Supreme Parish and Collegiate Church', 'literally', 'Supreme Parish and Cathedral Church', 'Berlin', 'Germany', 'Evangelical congregation Gemeinde', 'Oberpfarr- und Domkirche', 'Berlin', 'umbrella organisation Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia', 'Museum Island', and 'Mitte borough'. A status message indicates 'Full labeling in 4 s (162 words), language is EN'. Further down, there are two paragraphs of text with labels applied to specific words and phrases. At the bottom, there is a link to 'Download detailed results HTML editor (Not active: please log to your account)' and a footer containing copyright information and a Creative Commons license link.

Figure 11. Wikimeta demo

<sup>28</sup> <http://unlock.edina.ac.uk/texts/introduction>

<sup>29</sup> <http://www.geonames.org/>

<sup>30</sup> <http://www.wikimeta.com/>

We tested Wikimeta with the demo text in English and in French. The result of the English annotation is displayed in Figure 11. The tool identified most entities, although there are some discrepancies with the classifications. For example, Berlin is identified as place in two cases and in one case after a comma identified as organization. Also, St. Mary's Church is not recognized as one entity but treated separately.

### Zemanta

Zemanta<sup>31</sup> analyzes user-generated content (e.g. a blog post) using natural language processing and semantic search technology to suggest pictures, tags and links to related articles. Zemanta's core technology is a semantic recommendations engine. Zemanta indexes content from brands, marketers and agencies, as well as from bloggers into their semantic engine. The technology used is natural language processing and semantic search technology. This indexed content is then recommended to bloggers in order to spread the content over the web. Bloggers can use the recommended content to enrich their blogs. These targeted recommendations should result in better search ranking and increased SEO traffic for the content providers. (Zemanta, 2012)

Zemanta for bloggers can be installed as a browser plug-in if the blogger is using a blogging platform. If the blogger owns a server, he can also use the server-side plug-in. It operates on English text only, but it can perform named entity recognition in other languages. We tested the Zemanta Demo<sup>32</sup> with the English demo text. The demo did not work that well. Zemanta found the most important entities, leading to links and recommended tags that do represent the text. Yet, the related images only contain a few suitable images and mainly display other evangelical churches. The related articles show two articles related to Berlin (but not the Berlin Cathedral) and the other ones have absolutely nothing to do with the text (cf. Figure 12).

---

<sup>31</sup> <http://www.zemanta.com/>

<sup>32</sup> <http://www.zemanta.com/demo/>

## D7.4: Market study on technical options for semantic feature extraction

**Zemanta Demo**  
*Just paste any text into the box below. Then watch Zemanta go to work, generating the most relevant images, links, tags, and related content. It's that simple—and smart.*

Try inserting an image by simply clicking on it.

**Download**

Figure 12. Zemanta Demo

Zemanta would provide Europeana the opportunity to use the software in two ways. The goal could be that Zemanta recommends content from Europeana to bloggers and other users of the plug-ins but it may also be used in Europeana to suggest more content to the Europeana user. At the moment Zemanta mainly recommends content from Wikipedia, YouTube, IMBD.com, Amazon, Twitter, etc. (Zemanta, 2012)

D7.4: Market study on technical options for semantic feature extraction

Feature Matrix

	DBpedia Spotlight	sw2sws	Open Calais	Zemanta	CiceroLite (formerly Extractiv)	PoolParty (Extractor)	Wikimeta	Luxid®	Alchemy API	Evri	Apache Stanbol (formerly FISE)	Sophia Semantic Engine	Ontonaut	Unlock Text	Open Sahara
<b>operates on</b>															
text	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
<b>License / business model*</b>															
commercial				x	x	x		x				x			
commercial but with academic options			x						x	x					
free with quota							x						x		
open source	x	x									x			x	x
<b>Installation</b>															
web service	x		x	x						x					x
local installation	x	x		x	x	x		x			x	x			
API	x		x	x	x		x		x	x	x		x	x	x
<b>technologies in use</b>															
NLP	x		x	x					x		x	x			
machine learning			x		x				x						
semantic search technology				x									x		
other technologies*			x			x			x		x		x	x	x
<b>based on</b>															
thesaurus(i)						x		x			x				
dictionary(ies)						x		x				x			
ontology(ies)	x	x	x								x				
DBpedia	x			x			x				x		x		
Linked Data or other web content*				x			x		x	x	x		x	x	

D7.4: Market study on technical options for semantic feature extraction

	DBpedia Spotlight	sw2sws	Open Calais	Zemanta	CiceroLite (formerly Extractiv)	PoolParty (Extractor)	Wikimeta	Luxid®	Alchemy API	Evri	Apache Stanbol (formerly FISE)	Sophia Semantic Engine	Ontonaut	Unlock Text	Open Sahara
<b>Multilingual capabilities</b>															
Automatic language recognition			x	x	x				x		x		x		
Supported languages*	en	es	en, fr, es	en	en, ar, zh	en, de	en, fr	en, fr, de, gr, hu, it, cz, nl, pl, pt, ru, es	en, fr, de, it, pt, ru, es, sv	en	en	en (+ others)	de, en	?!	en, nl
Cross-language capabilities															
<b>Named Entities extraction</b>															
Noun groups extraction	x	x	x		x	x	x	x	x	x	x	x	x	x	x
Related terms / tags suggestions	x		x	x		x		x		x	x				
Relation extraction			x		x				x	x					
Classification (event, activity, name, person, etc.)	x		x		x	x	x	x	x	x		x			x
Confidence or relevance calculation	x		x												
Sentence, paragraph, document structure identification									x			x			
Related documents, pictures, links				x		x	x	x		x	x		x		
Identification of overall topics			x		x										

\* further explanation in the text

## Tools for semantic extraction – multimedia

This chapter lists all tools in alphabetical order that perform semantic extraction on multimedia documents matching our requirements defined in the chapter *Scope of this study*.

### Annnotation and SugarTube

Annnotation<sup>33</sup> and SugarTube<sup>34</sup> are two tools developed by the Open University (Milton Keynes, UK) that allow annotating videos with Linked Data resources, subsequently navigating them and enriching them with additional materials. Figure 13 displays Annnotation's user interface. It shows the actual video, the added tags and the possibility to search for suitable Linked Data URIs to annotate the video. (Lambert & Yu, 2010)

There is no automatic or semi-automatic entity extraction implemented.

The screenshot displays the Annnotation web interface. At the top, the KMi logo and the title 'Annnotation' are visible, along with a message: 'You cannot make annotations until you login'. The video title is 'Woods/ East Berlin'. The interface includes a video player on the left, a list of tags on the right, a mood section, and a geographical map of Berlin with a list of tags for different locations.

Time	Duration	Tag
00:00	01:56	<a href="http://dbpedia.org/resource/Berlin">http://dbpedia.org/resource/Berlin</a>
00:00	01:56	<a href="http://dewey.info/class/943/">http://dewey.info/class/943/</a>
00:00	01:56	<a href="http://id.loc.gov/authorities/sh85013353">http://id.loc.gov/authorities/sh85013353</a>
00:00	01:56	<a href="http://dbpedia.org/resource/Aftermath_of_World_War_II">http://dbpedia.org/resource/Aftermath_of_World_War_II</a>
00:00	01:56	<a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a>
00:41		<a href="http://dbpedia.org/resource/Vuvuzela">http://dbpedia.org/resource/Vuvuzela</a>
00:45	00:20	<a href="http://dbpedia.org/resource/Hoarding">http://dbpedia.org/resource/Hoarding</a>
00:58		<a href="http://dbpedia.org/resource/Tank">http://dbpedia.org/resource/Tank</a>
01:27		<a href="http://dbpedia.org/resource/Doughnut">http://dbpedia.org/resource/Doughnut</a>

Use Show on Map	Tag
Map	<a href="#">Berlin, DE [City, town, village...]</a>
Map	<a href="#">Berlin, US [City, town, village...]</a>
Map	<a href="#">Berlin, US [City, town, village...]</a>
Map	<a href="#">Khmelevoye, UA [City, town, village...]</a>
Map	<a href="#">Bartind, SE [mountain, hill, rock...]</a>
Map	<a href="#">Berlin, VE [City, town, village...]</a>
Map	<a href="#">Berlin, CO [City, town, village...]</a>
Map	<a href="#">Berlin, CO [City, town, village...]</a>
Map	<a href="#">Berlin, CO [City, town, village...]</a>
Map	<a href="#">Berlin, CO [City, town, village...]</a>
Map	<a href="#">Berlin de Cortés, MN [City, town, village...]</a>

Figure 13. Annnotation Demo

SugarTube, on the other hand, is a browser to search for the annotated videos and explore related content through the Linked Data resources. The user can either type in keywords or find content through geographical maps, entity recognition of natural language text supplied directly or from a URI, or by time. (Lambert & Yu, 2010) Figure 14 shows SugarTube's architecture.

<sup>33</sup> <http://annnotation.open.ac.uk/annnotation>

<sup>34</sup> <http://notube.open.ac.uk/sugartube/index.html>



Figure 14. SugarTube Demo

### Automatic semantic video indexing of Turkish news videos

Küçük & Yazıcı (2011) describe in their paper a video annotation and retrieval system which performs automatic semantic annotation of news video archives. Unfortunately we could not find a demo online. The system works for Turkish news videos, yet the system is a generic one that tries to overcome the semantic gap between what can be automatically extracted and the actual human perception of the semantics.

The system exploits the video texts as information source, such as speech transcriptions, overlay/sliding texts, and webcast texts. These video texts can be obtained by automatic speech recognition (ASR), video optical character recognition, or sliding text recognition. The system also employs several information extraction techniques, such as: named entity recognition, person entity extraction, co-reference resolution, and semantic event extraction. Other employed technologies are news story segmentation, text extraction and video retrieval (having a news video database in the background). Figure 15 displays the workflow of the semantic video indexing. In particular, the steps of automatic hyperlinking (Web alignment) and event extraction are interesting features. Automatic hyperlinking refers to exploiting web articles that are published at the same time in order to find matching articles and extracting their semantics. Event extraction is mostly performed in sports video analysis as these contain reoccurring events that can be identified more easily. Therefore, this system can only provide event extraction for a specific set of events that is based on events detected during the keyword analysis and include: Statement, Death, Trial/Investigation, Crash, Weather, etc. (Küçük & Yazıcı, 2011)

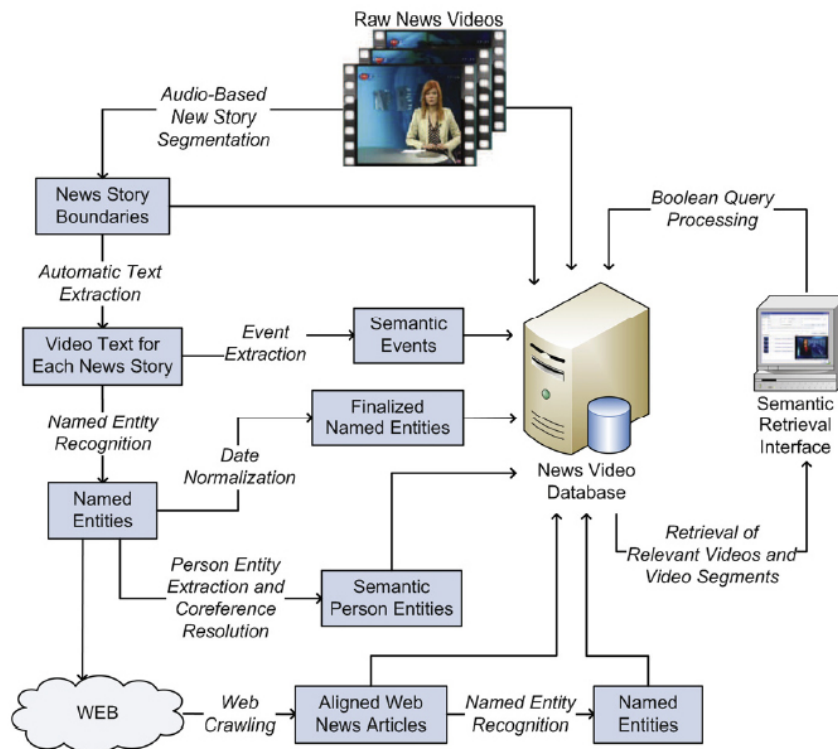


Figure 15. Semantic Video Indexing (Küçük &amp; Yazıcı, 2011)

## Contentus

Within the Contentus project<sup>35</sup> semantic analysis of video and audio sequences has been researched. In their article, Nandzik et al. (2012) describe the method of how video and audio documents can be semantically analyzed to be made fit for semantic information retrieval. Therefore the methods described are part of a larger framework to serve Contentus' semantic search and retrieval engine. Unfortunately we could not find a demonstrator online.

Video sequences are processed through three main steps: quality analysis, restoration and semantic analysis. The process of semantic analysis works as follows: the video sequences are temporally divided into shots and sub-shots and then the median picture of each sub-shot is used to represent the entire shot. All these key shots are then gathered and annotated with a bag-of-features based approach. Over 90 categories can be defined (e.g. indoor/outdoor, male/female) and images classified accordingly.

Audio sequences go through a similar process consisting of quality analysis, segmentation, speaker and speech recognition and musical analysis. We considered all steps except the quality analysis as part of the semantic analysis of audio documents. Firstly, audio sequences are segmented into speech, silence, and music. Also, it can be detected whether the audio sequence was recorded directly or over the phone (as this makes a difference for the automatic speech recognition) and the gender of the speaker can be detected. The automatic speech recognition system works on German speech and was trained on broadcast news data, pronunciation dictionaries with 200,000 words and 10,000 syllables as well as some specific language models. The system is also able to recognize specific speakers. The last step is the musical analysis where low-, mid- and high-level audio features from the MPEG-7 standard are recorded as well as musical attributes including

<sup>35</sup> <http://www.contentus-projekt.de/index.html?&L=1>



mood descriptions or other descriptive tags (e.g. synthetic, acoustic, etc.). To find similar music all available information is combined and a music ontology is used.

After these processes all necessary information is available in a textual form which then can be semantically analyzed using a similar tool described in the first part of the market study.

## Impala

Impala<sup>36</sup> is a tool that is employed by Beeld en Geluid. It was developed by Euvision Technologies<sup>37</sup> (a spin-off from the University of Amsterdam). Impala can detect concepts from images and videos. Impala adds visually-based tags to content automatically. The technology is based on machine-learning. That is why it basically can learn any concept if enough examples of the concept in a visual form are provided. Yet, this also means that the software cannot perform named entity recognition as it focuses on a more general recognition (e.g. age, gender, looks, etc.). For example, the system would identify a blonde-haired female in her thirties and not recognize Paris Hilton. Figure 16 shows how Impala found images depicting US flags.

A feature called “copy detection” is able to identify exact duplicates. With the help of “near-copy detection” the system can recognize resembling 3D objects or real world scenes (without storing them in a database first, like Amazon’s A9). Impala can identify concepts from the following categories: animals & plants, city life, faces, landscapes, man-made objects, people, sports and style. The specific concepts from each category are listed on their website<sup>38</sup>. (Euvision Technologies, 2012)

Euvision has the exclusive right to sublicense the Impala Core engines that were mainly developed at the University of Amsterdam.

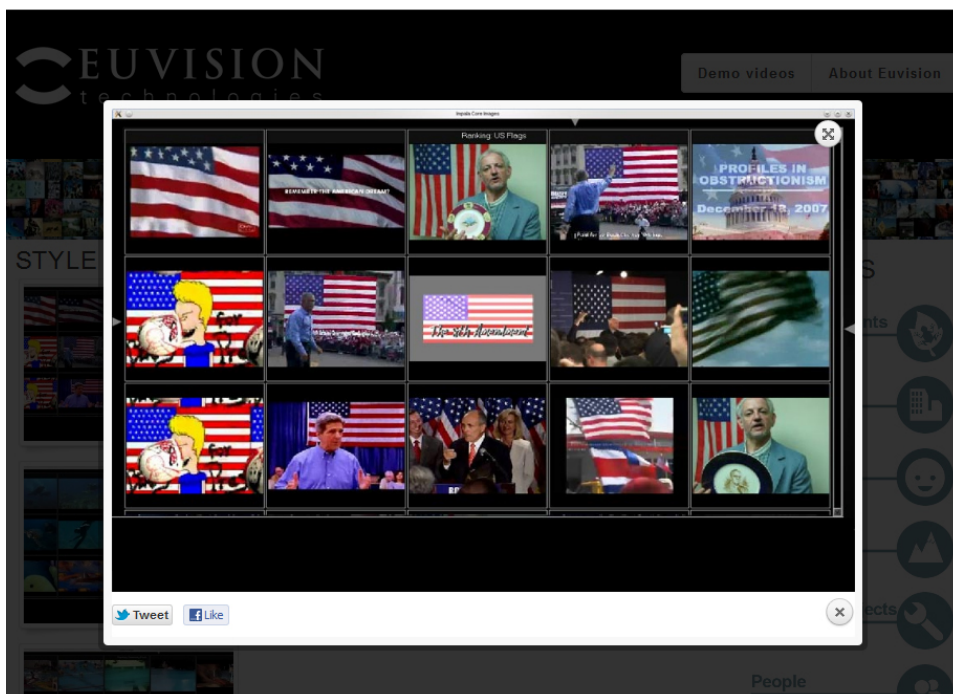


Figure 16. Impala Demo

<sup>36</sup> <http://www.euvt.eu/impala/>

<sup>37</sup> <http://www.euvt.eu>

<sup>38</sup> <http://www.euvt.eu/impala-concepts/>

### KAT (K-Space Annotation Tool)

KAT<sup>39</sup> is a framework developed under the lead of the University of Koblenz for semi-automatic, semantic annotation of images (Dasiopoulou, 2011). The tool is based on the Multimedia Metadata Ontology (M3O) which is a follow-up initiative of the Core Ontology of Multimedia (COMM) and provides descriptive and structural (according to MPEG-7 specifications) annotations. KAT is an open source tool, released under the terms of the LGPL. It consists of a graphical user interface, a plug-in infrastructure and a set of standard plug-ins for image annotation, retrieval, and browsing of ontologies. Input languages are RDFS and OWL; the produced annotations are in OWL. KAT is intended for combining automatic and manual annotation of images, which should improve the annotation process for the user. (Saathoff, 2009) Figure 17 displays a screenshot of the user interface.

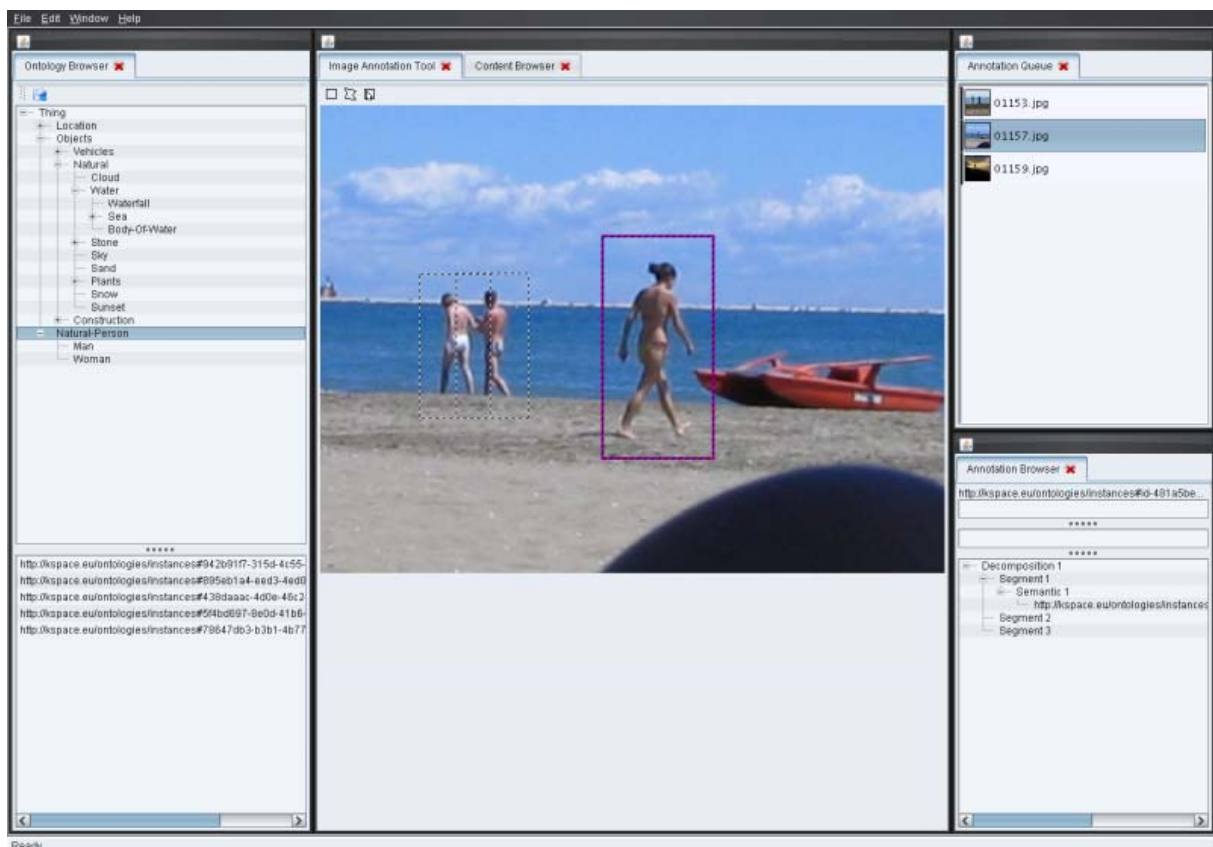


Figure 17. KAT Demo

### Semantic Video Annotation Suite

The Semantic Video Annotation Suite<sup>40</sup> (SVAS) was developed by Joanneum Research. It provides several automatic feature extraction methods that are not further specified on the website. It can perform object recognition and has a search tool which allows the recognition of specific objects throughout a video. (Joanneum Research)

The suite consists of two tools: the Media Analyzer, which extracts structural information automatically, and the Semantic Video Annotation Tool (SVAT), which allows editing the structural information from the Media Analyzer and adding further metadata according to

<sup>39</sup> <https://launchpad.net/kat>

<sup>40</sup> <http://www.joanneum.at/digital/produkte-loesungen/semantic-video-annotation.html>

#### D7.4: Market study on technical options for semantic feature extraction

MPEG-7. These metadata can either be administrative (creator, production date, title, etc.) or descriptive, capturing subject matters about persons, places, events, etc. These descriptive elements can be added to a shot (segment) or region level either manually or by using automatic image segmentation. Objects identified like this can then be used to detect similar objects throughout the video with the automatic matching service. The output format of the annotations is a MPEG-7 XML file. (Dasiopoulou, 2011)

There is a demo version available on the website which can be tested for 30 days after contacting Joanneum Research. Further use needs to be negotiated with Joanneum Research.

## References

- AlchemyAPI. <http://www.alchemyapi.com/> (03.02.2012)
- Apache Stanbol, <http://incubator.apache.org/stanbol/> (20.03.2012)
- Berlin Cathedral. (2011, December 21). In *Wikipedia, The Free Encyclopedia*. Retrieved 09:35, January 19, 2012, from [http://en.wikipedia.org/w/index.php?title=Berlin\\_Cathedral&oldid=467075553](http://en.wikipedia.org/w/index.php?title=Berlin_Cathedral&oldid=467075553)
- Berliner Dom. (2012, January 13) In *Wikipedia, Die freie Enzyklopädie*. Retrieved 09:38, January 19, 2012, from [http://de.wikipedia.org/w/index.php?title=Berliner\\_Dom&oldid=98811237](http://de.wikipedia.org/w/index.php?title=Berliner_Dom&oldid=98811237)
- Berliner Dom. (2012, février 1). *Wikipédia, l'encyclopédie libre*. Page consultée le 10:27, février 16, 2012 à partir de [http://fr.wikipedia.org/w/index.php?title=Berliner\\_Dom&oldid=75024851](http://fr.wikipedia.org/w/index.php?title=Berliner_Dom&oldid=75024851).
- Blumauer, A. (2012). Personal communication with Marlies Olensky.
- Catedral de Berlín. (2011, 28 de diciembre). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 10:28, febrero 16, 2012 desde [http://es.wikipedia.org/w/index.php?title=Catedral\\_de\\_Berl%C3%ADn&oldid=52529159](http://es.wikipedia.org/w/index.php?title=Catedral_de_Berl%C3%ADn&oldid=52529159).
- CELL. (2002). Sophia Semantic Engine. Retrieved from <http://www.celi.it/en/sophia-semantic-engine.shtml> (10.04.2012)
- Criado-Fernandez, L. (2010). MediaWiki de la herramienta Sw2sws. Retrieved from [http://sourceforge.net/userapps/mediawiki/lcriadof/index.php?title=Main\\_Page](http://sourceforge.net/userapps/mediawiki/lcriadof/index.php?title=Main_Page) (19.01.2012)
- Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris Y. (2011). A Survey of Semantic Image and Video Annotation Tools In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, et al. (Eds.), *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. Lecture Notes in Computer Science 6050* (pp. 196–239). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Euvison Technologies. (2012). Impala. Retrieved from <http://www.euvt.eu/impala/> (16.04.2012)
- Evri Developers, <http://corporate.evri.com/developers/> (10.04.2012)
- Evri Developer Center, <http://www.evri.com/developer> (10.04.2012)
- Freire, N., Mane, L., & Petz, G. (2008). State of the art of semantic and multilingual engines or tools for digital libraries. Deliverable D3.1, TELplus-Project.
- Joanneum Research. Semantic Video Annotation Suite. Retrieved from <http://www.joanneum.at/de/digital/produkte-loesungen/semantic-video-annotation.html> (28.03.2012)
- Küçük, D., & Yazıcı, A. (2011). Exploiting information extraction techniques for automatic semantic video indexing with an application to Turkish news videos. In *Knowledge-Based Systems*, Vol 42, pp 844-857.
- Lambert, D., & Yu, H.Q. (2010). Linked Data based video annotation and browsing for distance learning. In *SemHE '10: The Second International Workshop on Semantic Web Applications in Higher Education, 3 November 2010, Southampton, UK*.
- Language Computer. CiceroLite: Text Annotation, Entity Extraction. <http://www.languagecomputer.com/products/text-annotation/cicerolite.html> (25.01.2012)
- Luxid® for Content Enrichment (TEMIS). <http://www.temis.com> (03.02.2012)
- Mendes, P. (2012). Personal communication with Marlies Olensky.
- Mendes, P.N., Jakob, M., García-Silva, A. & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*. Graz, Austria, 7–9 September 2011. Retrieved from

- <http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Mendes-Jakob-GarciaSilva-Bizer-DBpediaSpotlight-ISEM2011.pdf> (19.01.2012)
- Mendes, P.N., Jakob, M., Daiber J. & Bizer, C. (2011). DBpedia Spotlight. <http://dbpedia.org/spotlight> (03.02.2012)
- Nandzik, J., Litz, B., Flores-Herr, N., Löhden, A., Konya, I., Baum, D., et al. (2012). CONTENTUS—technologies for next generation multimedia libraries. *Multimedia Tools and Applications*. Retrieved from <http://www.springerlink.com/content/b115174541hxg628/fulltext.pdf> (04.04.2012)
- OpenCalais. <http://www.opencalais.com/> (24.01.2012)
- Open Sahara, <http://opensahara.com/> (10.04.2012)
- Ontonaut, <http://ontonaut.net> (10.04.2012)
- Rizzo, G. & Troncy, R. (2011). NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *Workshop on Web Scale Knowledge Extraction (WEKEX'11), Vol 20*.
- Rizzo, G., Troncy, R., Hellmann, S. & Bruemmer, M. (2012). NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In *WWW2012 Workshop: Linked Data on the Web (LDOW2012)*. Retrieved from <http://events.linkedata.org/ldow2012/papers/ldow2012-paper-02.pdf> (27.04.2012)
- Saathoff, C. K-Space Annotation Tool. Retrieved from <http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/systeme/kat> (04.04.2012)
- SciVal: Elsevier Fingerprint Engine™. (2012). <http://www.info.scival.com/experts>
- Semantic Web Company. (2012). PoolParty: Semantic Information Management. <http://www.semantic-web.at/poolparty-semantic-information-management> (26.01.2012)
- Sw2sws. <http://sw2sws.sourceforge.net/> (19.01.2012)
- Unlock. (2012). Unlock Text. Retrieved from <http://unlock.edina.ac.uk/home/> (10.04.2012)
- Walsh, J. (2012). Personal communication with Marlies Olensky.
- Wikimeta. <http://www.wikimeta.com/> (02.02.2012)
- Worrying, M. (2008). Semantic Video Indexing. 2008 Video Search Summit. Retrieved from <http://www.reelseo.com/semantic-video-indexing/> (13.04.2012)
- Zemanta. <http://www.zemanta.com/> (25.01.2012)
- Zemanta. (2012, February 5). In *Wikipedia, The Free Encyclopedia*. Retrieved 16:30, February 14, 2012, from <http://en.wikipedia.org/w/index.php?title=Zemanta&oldid=475205227>

## Appendix

### Demo texts

These are the texts that were used to shortly test demo versions of the available tools. The texts were simply copied from Wikipedia.

#### English

Berlin Cathedral (German: Berliner Dom) is the colloquial name for the Evangelical (i.e. Protestant) Oberpfarr- und Domkirche (English analogously: Supreme Parish and Collegiate Church, literally Supreme Parish and Cathedral Church) in Berlin, Germany. It is the parish church of the Evangelical congregation Gemeinde der Oberpfarr- und Domkirche zu Berlin, a member of the umbrella organisation Evangelical Church of Berlin-Brandenburg-Silesian Upper Lusatia. Its present building is located on Museum Island in the Mitte borough. The Berlin Cathedral had never been a cathedral in the actual sense of that term since it has never been the seat of a bishop. The bishop of the Evangelical Church in Berlin-Brandenburg (under this name 1945–2003) is based in St. Mary's Church, Berlin, and Kaiser Wilhelm Memorial Church. St. Hedwig's Cathedral serves as seat of Berlin's Roman Catholic metropolitan bishop.

Berlin Cathedral. (2011, December 21). In *Wikipedia, The Free Encyclopedia*. Retrieved 09:35, January 19, 2012, from [http://en.wikipedia.org/w/index.php?title=Berlin\\_Cathedral&oldid=467075553](http://en.wikipedia.org/w/index.php?title=Berlin_Cathedral&oldid=467075553)

What entities or concepts could be detected?

Berlin and Germany – geographical entities

German – language

Berlin Cathedral, Berliner Dom – named entity / translations

Translations: Oberpfarr- und Domkirche, Supreme Parish and Collegiate Church, Supreme Parish and Cathedral Church

Evangelical congregation - concept

umbrella organisation – concept

Berlin-Brandenburg-Silesian Upper Lusatia – geographical entities

Museum-Island – geographical entity

Mitte borough – geographical entity

seat of a bishop – concept

St. Mary's Church – named entity

Kaiser Wilhelm Memorial Church – named entity

St. Hedwig's Cathedral - named entity

Roman Catholic metropolitan bishop – concept

1945–2003 – time-span

#### German

Der Berliner Dom (eigentlich Oberpfarr- und Domkirche zu Berlin) ist eine evangelische Kirche im Berliner Ortsteil Mitte des gleichnamigen Bezirks auf dem nördlichen Teil der Spreeinsel, die hier Museumsinsel genannt wird. Der 1894 bis 1905 nach Plänen von Julius Raschdorff in Anlehnung an die italienische Hochrenaissance und den Barock errichtete Dom gehört zu den bedeutendsten protestantischen Kirchenbauten in Deutschland. Das denkmalgeschützte Gebäude besteht aus der zentralen Predigtkirche unter der Kuppel sowie der Tauf- und Trau Kirche. Das Hauptportal liegt am Lustgarten. In der Gruft des Doms ruhen zahlreiche Mitglieder des Hauses Hohenzollern. Die Kuppelkonstruktion wurde 2007 für die Auszeichnung als Historisches Wahrzeichen der Ingenieurbaukunst in Deutschland nominiert. Die Gesamthöhe beträgt heute 116 Meter auf einer Grundfläche von 114 x 73 Meter. Die Kuppel besitzt eine Scheitelhöhe von 74,8 Meter bei einem Durchmesser von 33 Meter. Heute finden im Berliner Dom neben den regelmäßigen Gemeindegottesdiensten

auch Gottesdienste anlässlich von Staatsakten oder wichtigen politischen Ereignissen der Bundesrepublik Deutschland statt.

Berliner Dom. (2012, January 13) In Wikipedia, Die freie Enzyklopädie. Retrieved 09:38, January 19, 2012, from

[http://de.wikipedia.org/w/index.php?title=Berliner\\_Dom&oldid=98811237](http://de.wikipedia.org/w/index.php?title=Berliner_Dom&oldid=98811237)

What entities or concepts could be detected?

Berliner Dom - named entity

Oberpfarr- und Domkirche zu Berlin - named entity

Berlin, Deutschland – geographical entities

evangelische Kirche – concept

Berliner Ortsteil Mitte – geographical entity

Spreeinsel, Museumsinsel – geographical entities

1894 bis 1905 – time-span

Julius Raschdorff – person

italienische Hochrenaissance – time-span/event

Barock – time-span/event

protestantischen Kirchenbauten – concept

Predigtkirche – concept

Tauf- und Traukirche – concept

Lustgarten – geographical entity

Hohenzollern – person, concept

Historisches Wahrzeichen – concept

French

Le Berliner Dom est la cathédrale historique de Berlin située sur l'île aux Musées. L'empereur Guillaume II voulant pour l'Église Luthérienne une cathédrale digne de la grandeur de la capitale impériale, il la fit construire sur le lieu d'une ancienne cathédrale du XVIIIe siècle. L'architecte Julius Carl Raschdorff en fut le maître d'œuvre. Sa construction s'étala entre 1894 et 1905. L'édifice de style Haute Renaissance italienne a une longueur de 114 mètres, pour une largeur de 73 mètres et une hauteur de 85 mètres. Le corps central est coiffé d'une imposante coupole, consacré au prêche. L'aile Sud comporte la chapelle des Baptêmes et des Mariages, tandis que l'aile Nord abrite la chapelle funéraire. La crypte réunit près de 95 sarcophages, où reposent les membres de la dynastie des Hohenzollern.

Berliner Dom. (2012, février 1). *Wikipédia, l'encyclopédie libre*. Page consultée le 10:27, février 16, 2012 à partir de

[http://fr.wikipedia.org/w/index.php?title=Berliner\\_Dom&oldid=75024851](http://fr.wikipedia.org/w/index.php?title=Berliner_Dom&oldid=75024851).

Spanish

La Catedral de Berlín (Berliner Dom en alemán) es un templo de la Iglesia Evangélica en Alemania ubicado en Berlín, Alemania. Cuando en 1930 la Santa Sede estableció por primera vez una diócesis católica en Berlín, la catedral de Berlín había sido ya un templo protestante por mucho tiempo. La catedral de Santa Eduvigis es el sitio de residencia del Obispo metropolitano de Berlín. El edificio fue construido entre 1895 y 1905. El lugar donde se encuentra este edificio lo ocupaba anteriormente una catedral barroca construida por Johann Boumann culminada en 1747 y posteriormente remodelada en 1822 por el arquitecto berlinés Karl Friedrich Schinkel en estilo neoclásico. Esta catedral fue demolida en 1894 por orden del emperador Guillermo II y fue reemplazada por la actual, diseñada por Julius Raschdorff en el estilo neobarroco de fines de S. XIX e inicios de S.XX. Durante la Segunda Guerra Mundial, el templo fue seriamente dañado por los bombardeos. Hasta 1975, fecha en la que comenzaron los trabajos de reconstrucción, se colocó un techo provisional para proteger el interior.

Catedral de Berlín. (2011, 28 de diciembre). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 10:28, febrero 16, 2012 desde

[http://es.wikipedia.org/w/index.php?title=Catedral\\_de\\_Berl%C3%ADn&oldid=52529159](http://es.wikipedia.org/w/index.php?title=Catedral_de_Berl%C3%ADn&oldid=52529159).

## Resources

- DiCiuccio, R. (2010). Blog on: Entity Extraction & Content API Evaluation. Retrieved from <http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation> (28.03.2012)
- Diefenthal, C. (2010). Blog on: entity extraction comparison tool – opencalais, alchemyAPI, evri. Retrieved from <http://veeeb.de/blog/news/entity-extraction-alchemyapi-evri-opencalais/> (21.03.2012)
- Fagan, M. (2011). Entity Extraction APIs, once again. Retrieved from <http://faganm.com/blog/2011/07/26/1014/> (26.03.2012)
- Fagan, M. (2010). Comparing NLP APIs for Entity Extraction. Retrieved from <http://faganm.com/blog/2010/01/02/1009/> (26.03.2012)
- Quénot, G., & Awad, G. (2011). Semantic Indexing task: Overview. *TRECVID-2011*. Retrieved from <http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.sin.slides.pdf> (04.04.2012)



Other exemplary state-of-the-art tools

GATE, a general architecture for text engineering

<http://gate.ac.uk/>

MALLET, Machine Learning for Language Toolkit

<http://mallet.cs.umass.edu/>

Stanford Named Entity Recognizer

<http://nlp.stanford.edu/software/CRF-NER.shtml>

Headup Entity Extraction

<http://labs.headup.com/Services/RealTime/API/EntityExtraction/Playground.aspx>

BANNER Named Entity Recognition System

<http://banner.sourceforge.net/>

Rosette Entity Extractor (REX)

<http://www.basistech.com/entity-extractor/>

AFNER - Named Entity Recognition

<http://afner.sourceforge.net/>