Semantically Enriching Content Using OpenCalais

Marius-Gabriel BUTUC

***Abstract*—One of the key challenges of the Semantic Web is how to go from today's unstructured web to a web rich with semantic information. Following the bottom up approach, OpenCalais is an automated system intended to our annotate data and information, allowing us to gradually build semantically enabled systems. By semantically enriching the published content, we help users enjoy their on-line experiences and reduce the frustration of dealing with voluminous amounts of information that is incoherently organized and often irrelevant to a particular person's need.**

***Index Terms*—API, Linked Data, resource management, web service.**

I. Introduction

CALAIS identifies and automatically tags entities (e.g., people, places, companies, geographies), facts (e.g., relationships) and events (e.g., things that happened) in text. It then fabricates connections between those entities and significant data sets, media files, and similar entries using open data from sources like Wikipedia, DBpedia, GeoNames, the Internet Movie Database (IMDB), Shopping.com, Reuters.com, etc.

With the original goal - help developers, bloggers and publishers automatically tag their content to improve search and navigation, and enable new reader engagement features - the team behind Calais has developed a tool to automate time consuming content operations and increase productivity.

This paper will look into the possibilities of semantically enriching published content [1] using Calais.

II. Calais and the OpenCalais Web Service

Calais is described by its developers as a "big initiative with a lot of components" [2], yet at its core lays the OpenCalais web service. The 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.

An overview of the Calais applications and tools is depicted in Fig. 1.

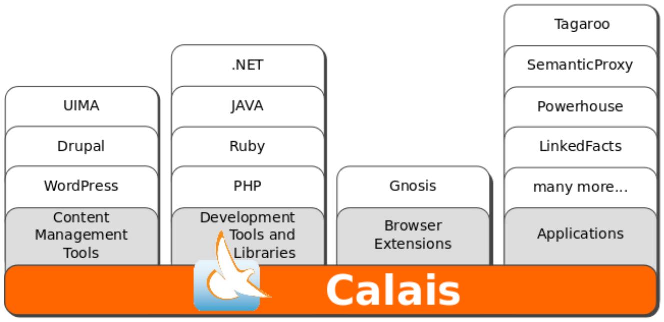


Fig. 1. Overview of Calais applications and tools.

Recently nominated one of the Top 10 Semantic Web products of 2009 [3], OpenCalais uses semantic technology and natural language processing to analyze text and add metadata by drawing out entities from documents, blog posts, news stories, etc. In some cases [4], this type of data can identify or help identify relationships between people, businesses, etc.

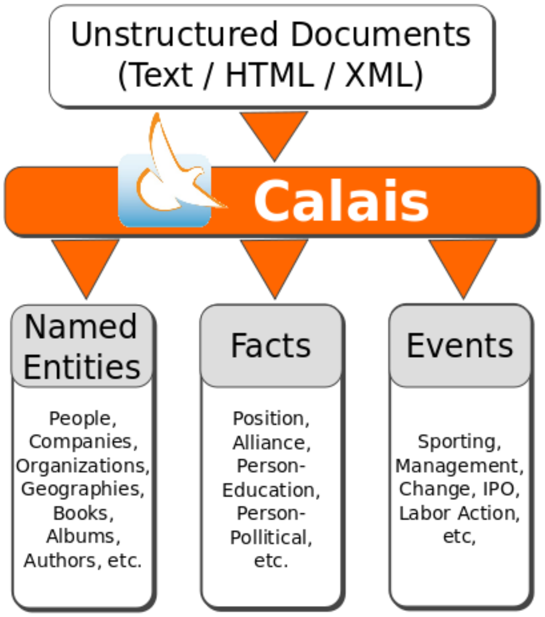

Unstructured Documents (Text / HTML / XML)
Calais
Named Entities
People, Companies, Organizations, Geographies, Books, Albums, Authors, etc.
Facts
Position, Alliance, Person-Education, Person-Pollitical, etc.
Events
Sporting, Management, Change, IPO, Labor Action, etc,

Fig. 2. OpenCalais web service.

A. How does it work?

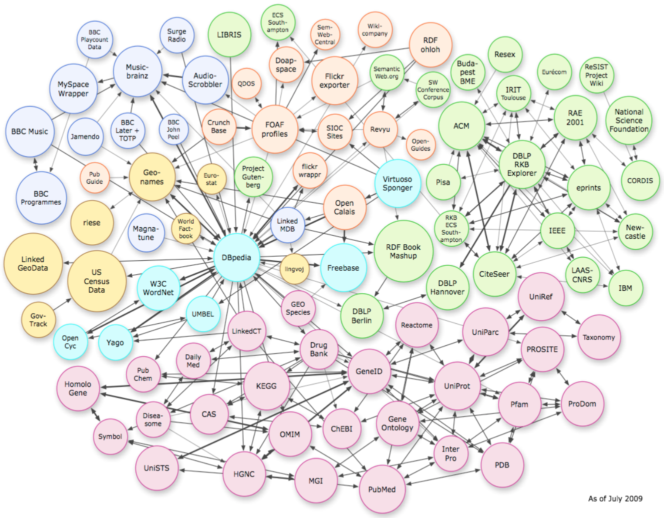
The basic interaction delivered by the OpenCalais API (Figure 2) can be summed up in the following steps:

Marius-Gabriel BUTUC is with the Department of Computer Science, Alexandru Ioan Cuza University, Iasi, Romania (e-mail: mbutuc@info.uaic.ro).

1) First of all, we need an API key – a string that uniquely identifies the specific instances of the application that uses the service.
2) After obtaining the key, we can submit content to the Calais Web service. The methods supported by the OpenCalais Web Service API are: SOAP, REST and HTTP Traffic Compression.
3) Calais tags each person, place, fact and event in the content, making it machine-readable and interoperable on the Web.
4) Each piece of content – and each entity or event in that content – is assigned a unique identifier (a document ID and many URIs) that ties back to the Linked Data Cloud.
5) We can use the metadata returned by Calais (tags, document IDs and URIs) in our application, e.g., we can use OpenCalais to scan a set of Web pages and build a local RDF store for querying and display.

B. *LinkedData Entities*

As of March 2009, Calais is oficially part of the Linked Open Data (LOD) Cloud (Figure 3).



As of July 2009

Fig. 3. The Linked Open Data Cloud – July 2009 [5].

Linked Data is a paradigm of exposing, sharing, and connecting data via dereferenceable URIs (Uniform Resource Identifier) on the Web. The method includes 4 principles [6]:

1) Use URIs to identify things that you expose to the Web as resources.
2) Use HTTP URIs so that people/machines can locate and dereference these things.
3) Provide useful information about the resource when its URI is dereferenced.
4) Include links to other, related URIs in the exposed data as a means of improving information discovery on the Web.

The Calais initiative has exposed its data via Linked Data endpoints. When Calais extracts an entity from a given text it also returns a dereferenceable entity URI. All the naming conventions for entities, facts, events and their corresponding attributes are rationalized and the schema is published on the official site as RDFS.

There are several classes of entities that are considered meaningful to Calais and they include `Currency`, `MarketIndex`, `Movie`, `MusicGroup`, `MusicAlbum`, `OperatingSystem`, `ProgrammingLanguage` and many others. The breadth and depth of information provided in each URI varies based on how well Calais can unambiguously identify this entity, automatically using its specific NLP algorithms.

As an example, "Calais" is an ambiguous name: we have both Calais – the semantic tool and Calais – the city. Even as a city name Calais is ambiguous, because there is one in the North of France and several in the USA. The URI of the ambiguous city "Calais" is:
`http://d.opencalais.com/genericHasher-1/82b41c38-0b5e-301f-8a84-839d4d78e78e.html`
If we try to dereference this URI, we'll get disambiguation options, each leading to another resource known to the API that can disambiguate "Calais". If we follow any of those resources, we will find useful information such as geographical coordinates and links to other Linked Data or Web assets.

Currently the list of *disambiguated* entities is limited to `Company`, `Product (Electronics)` and Geographies: `City`, `Country` and `ProvinceOrState`. `Company`, for example, is the richest of endpoints providing rich information, such as ticker symbol, officers and directors, corporate website, industry codes, etc.

C. *Interpreting the API Response*

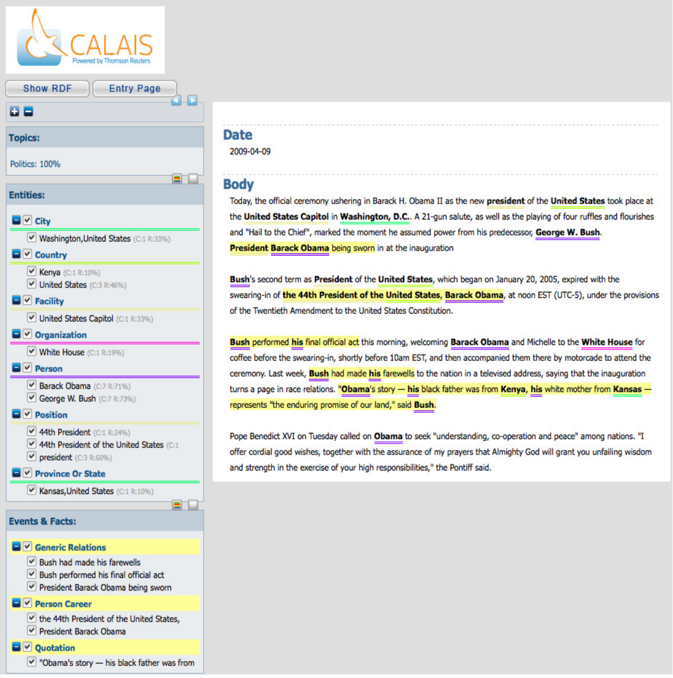
The Calais response can be easily understood [7] from RDF, JSON, microformats [8], [9] – a web-based approach to semantic markup that seeks to re-use existing (X)HTML tags to convey metadata and other attributes – or a simple HTML format. Another example of the implementation of the LOD principles is given when we look into the RDF representation for Moscow, Russia. This entity is linked via `owl:sameAs` to

- `http://dbpedia.org/resource/Moscow`
- `http://sws.geonames.org/524901/`
- `http://rdf.freebase.co/ns/guid.9202a8c04000641f800000000002636c`

III. Examples of applications that use OpenCalais

A. *Calais Viewer*

Calais Viewer [10] 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 Wikinews teaser about Barack Obama and got the results depicted in Figure 4. It identifies the correct topic of the article – `Politics`, the `City` that the article is about – `Washington, United States`, two `Person` entities: `Barack Obama` and `George W. Bush` etc.



Topics:

Politics: 100%

Entities:

- ☑ City
  - ☑ Washington,United States (C:1 R:33%)
- ☑ Country
  - ☑ Kenya (C:1 R:10%)
  - ☑ United States (C:3 R:46%)
- ☑ Facility
  - ☑ United States Capitol (C:1 R:33%)
- ☑ Organization
  - ☑ White House (C:1 R:19%)
- ☑ Person
  - ☑ Barack Obama (C:7 R:71%)
  - ☑ George W. Bush (C:7 R:73%)
- ☑ Position
  - ☑ 44th President (C:1 R:24%)
  - ☑ 44th President of the United States (C:1
  - ☑ president (C:3 R:60%)
- ☑ Province Or State
  - ☑ Kansas,United States (C:1 R:10%)

Events & Facts:

- ☑ Generic Relations
  - ☑ Bush had made his farewells
  - ☑ Bush performed his final official act
  - ☑ President Barack Obama being sworn
- ☑ Person Career
  - ☑ the 44th President of the United States,
  - ☑ President Barack Obama
- ☑ Quotation
  - ☑ "Obama's story — his black father was from

Date

2009-04-09

Body

Today, the official ceremony ushering in Barack H. Obama II as the new **president** of the **United States** took place at the **United States Capitol** in **Washington, D.C.**. A 21-gun salute, as well as the playing of four ruffles and flourishes and "Hail to the Chief", marked the moment he assumed power from his predecessor, **George W. Bush**.
President Barack Obama being sworn in at the inauguration

Bush's second term as **President** of the **United States**, which began on January 20, 2005, expired with the swearing-in of **the 44th President of the United States**, **Barack Obama**, at noon EST (UTC-5), under the provisions of the Twentieth Amendment to the United States Constitution.

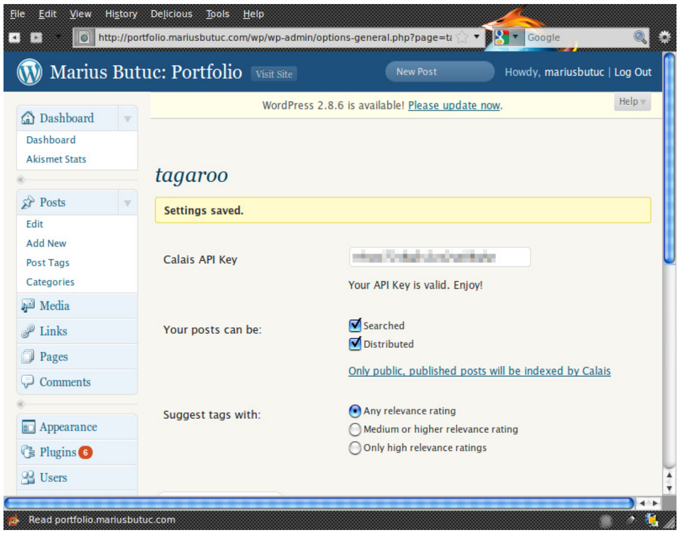
Bush performed **his** final official act this morning, welcoming **Barack Obama** and Michelle to the **White House** for coffee before the swearing-in, shortly before 10am EST, and then accompanied them there by motorcade to attend the ceremony. Last week, **Bush** had made **his** farewells to the nation in a televised address, saying that the inauguration turns a page in race relations. "**Obama**'s story — **his** black father was from **Kenya**, **his** white mother from **Kansas** — represents "the enduring promise of our land," said **Bush**.

Pope Benedict XVI on Tuesday called on **Obama** to seek "understanding, co-operation and peace" among nations. "I offer cordial good wishes, together with the assurance of my prayers that Almighty God will grant you unfailing wisdom and strength in the exercise of your high responsibilities," the Pontiff said.

Fig. 4. Calais Viewer in action: semantically enriching a Wikinews teaser.

B. Tagaroo

As we saw in Figure 1, Calais provides Content Management Tools for Drupal and Wordpress. Tagaroo [11] is one of those tools aimed to make any WordPress blog better for the publisher, for the readers and more accessible to search engines. Similar to Zemanta [12], Tagaroo analyzes the text in our post and suggests intelligent tags for the things and events we're writing about. Before using it, we need to have a valid API key (Figure 5).



http://portfolio.mariusbutuc.com/wp/wp-admin/options-general.php?page=t

WordPress 2.8.6 is available! Please update now.

tagaroo

Settings saved.

Calais API Key [redacted]

Your API Key is valid. Enjoy!

Your posts can be: ☑ Searched ☑ Distributed

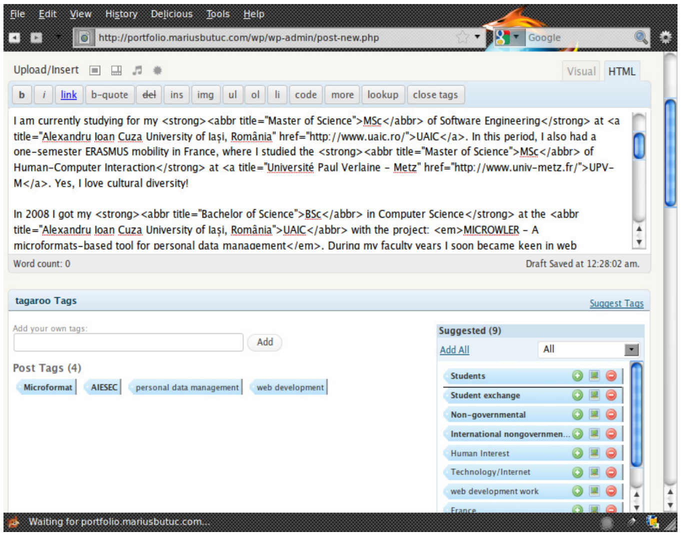
Only public, published posts will be indexed by Calais

Suggest tags with: ● Any relevance rating ○ Medium or higher relevance rating ○ Only high relevance ratings

Read portfolio.mariusbutuc.com

Fig. 5. Installing Tagaroo.

The user interface is intuitive and seamlessly integrated in WordPress. In order to semantically enrich our content, all we need to do is to start writing our post. After writing at least 64 characters, Tagaroo searches for tags. We can add them to our post, ignore them, look up Flickr images for any of them as illustrated in Figure 6 or even add our own tags.



http://portfolio.mariusbutuc.com/wp/wp-admin/post-new.php

```
I am currently studying for my <strong><abbr title="Master of Science">MSc</abbr> of Software Engineering</strong> at <a
title="Alexandru Ioan Cuza University of Iași, România" href="http://www.uaic.ro/">UAIC</a>. In this period, I also had a
one-semester ERASMUS mobility in France, where I studied the <strong><abbr title="Master of Science">MSc</abbr> of
Human-Computer Interaction</strong> at <a title="Université Paul Verlaine - Metz" href="http://www.univ-metz.fr/">UPV-
M</a>. Yes, I love cultural diversity!

In 2008 I got my <strong><abbr title="Bachelor of Science">BSc</abbr> in Computer Science</strong> at the <abbr
title="Alexandru Ioan Cuza University of Iași, România">UAIC</abbr> with the project: <em>MICROWLER - A
microformats-based tool for personal data management</em>. During my faculty years I soon became keen in web
```

Word count: 0

Draft Saved at 12:28:02 am.

tagaroo Tags

Suggest Tags

Add your own tags:

Post Tags (4)

Microformat | AIESEC | personal data management | web development

Suggested (9)

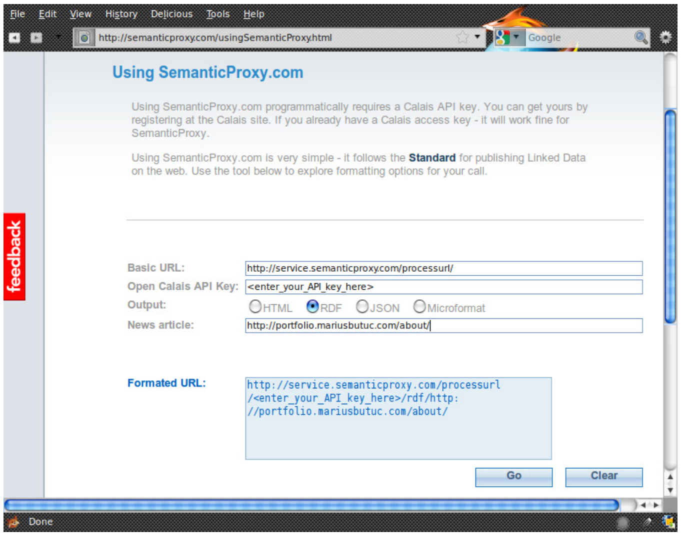
- Students
- Student exchange
- Non-governmental
- International nongovernmen...
- Human Interest
- Technology/Internet
- web development work
- France

Waiting for portfolio.mariusbutuc.com...

Fig. 6. Tagaroo in action: semantically enriching our content as we type.

C. SemanticProxy

Following the standard for publishing Linked Data on the Web [13], the Calais team developed SemanticProxy [14] to translate the content of any URL on the web to its semantic representation in RDF, JSON, microformats or HTML (Figure 8). It is optimized for performance on 30 of the world's largest English-language news sites, but it also works well with other sites like in Figure 7. Intended for the "little semantic machines love to come by and spend a few highquality milliseconds" [14], it is meant to leverage both current and future semantic applications.



http://semanticproxy.com/usingSemanticProxy.html

Using SemanticProxy.com

Using SemanticProxy.com programmatically requires a Calais API key. You can get yours by registering at the Calais site. If you already have a Calais access key - it will work fine for SemanticProxy.

Using SemanticProxy.com is very simple - it follows the **Standard** for publishing Linked Data on the web. Use the tool below to explore formatting options for your call.

Basic URL: http://service.semanticproxy.com/processurl/

Open Calais API Key: <enter_your_API_key_here>

Output: ○ HTML ● RDF ○ JSON ○ Microformat

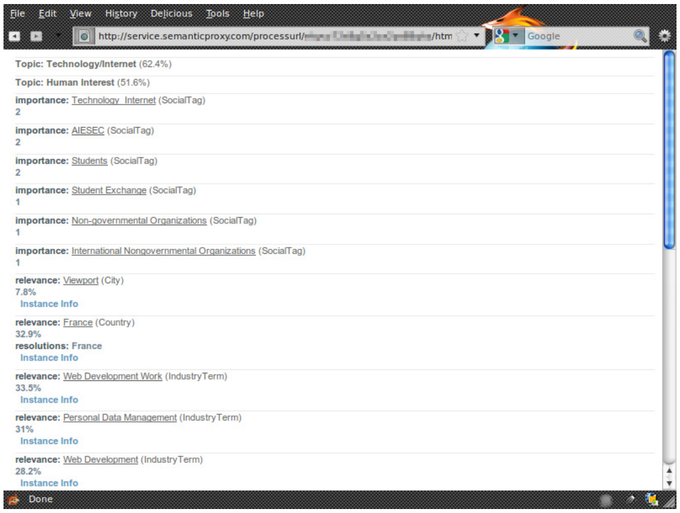
News article: http://portfolio.mariusbutuc.com/about/

Formated URL:

```
http://service.semanticproxy.com/processurl
/<enter_your_API_key_here>/rdf/http:
//portfolio.mariusbutuc.com/about/
```

Done

Fig. 7. SemanticProxy: Interface for the human user.



http://service.semanticproxy.com/processurl/[redacted]/htm

Topic: Technology/Internet (62.4%)

Topic: Human Interest (51.6%)

importance: Technology_Internet (SocialTag)
2

importance: AIESEC (SocialTag)
2

importance: Students (SocialTag)
2

importance: Student Exchange (SocialTag)
1

importance: Non-governmental Organizations (SocialTag)
1

importance: International Nongovernmental Organizations (SocialTag)
1

relevance: Viewport (City)
7.8%
Instance Info

relevance: France (Country)
32.9%
resolutions: France
Instance Info

relevance: Web Development Work (IndustryTerm)
33.5%
Instance Info

relevance: Personal Data Management (IndustryTerm)
31%
Instance Info

relevance: Web Development (IndustryTerm)
28.2%
Instance Info

Fig. 8. SemanticProxy in action: semantically enriched HTML results.

IV. Conclusion

OpenCalais currently only supports content in English, French and Spanish and it tries to identify automatically among the given languages by applying a Language Identification module before processing the text for entities, events and facts. Supporting more languages could be a good direction for development, opening Calais even more. The web service is free for both commercial and non-commercial use. The current limitations are 50,000 transactions per day with a frequency of maximum 4 transactions per second while an average transaction is expected to take from 0.5 to 1 second.

Using a semantic-based tool, the users are able to easily automate content operations, increase productivity and cut costs. Since the Semantic Web goal is not to build a web of data, but to enrich lives through access to information [15], we can now approach the bottom up paradigm [16], enhancing the value of our content and improving the user experience.

As stated in Section II-A, we can use OpenCalais, or even better SemanticProxy, to scan a set of Web pages and build a local RDF store for querying and display.

Another application could consist in a visualization tool that could make OpenCalais even more powerful. For example it might be interesting for visualization tools like Muckety [17] or NNDB Mapper [18] and to quickly ”see” relationships that might go unnoticed without tools like OpenCalais.

We intend to experiment the use of OpenCalais for enhancing the content of the academic social Web applications (e.g., blogs, wikis) in the context of the semantic e-learning.

References

[1] D. Allemang and J. Hendler, Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann, 2008.

[2] K. Thomas, Simple OpenCalais Whitepaper, http://slidesha.re/qoPWP, 2009.

[3] R. MacManus, Top 10 Semantic Web Products of 2009, http://www.readwriteweb.com/archives/top_10_semantic_web_products_of_2009.php, 2009.

[4] T. Segaran, C. Evans, and J. Taylor, Programming the Semantic Web. O'Reilly Media, 2009.

[5] ***, Linked Data - Connect Distributed Data across the web, http://linkeddata.org/.

[6] S. Buraga, Arhitectura aplicatiilor Web-ului semantic. Linked Data, Alexandru Ioan Cuza University Iasi, Tech. Rep., 2009, http://profs.info.uaic.ro/_busaco/teach/courses/wade/presentations/web08ArhitecturaAplicatiilorWebSemantic-LinkedData.pdf.

[7] ***, "Understanding the OpenCalais RDF Response," http://blog.3kbo.com/2009/09/26/opencalais-response/.

[8] ****, "Microformats," http://microformats.org/.

[9] J. Allsopp, Microformats: Empowering Your Markup for Web 2.0. Friends of Ed, 2007.

[10] ***, "Calais viewer," http://viewer.opencalais.com/.

[11] ***, "Tagaroo – make blogging better!" http://tagaroo.opencalais.com/.

[12] ***, "Zemanta: Blog smarter," http://www.zemanta.com/.

[13] C. Bizer, R. Cyganiak, and T. Heath, How to Publish Linked Data on the Web, Freie Universitaet Berlin, Tech. Rep., 2007, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

[14] ***, "Semanticproxy," http://semanticproxy.com/.

[15] I. Davis, If you love something... set it free, http://slidesha.re/haCax, 2009.

[16] A. Iskold, Semantic Web Patterns: A Guide to Semantic Technologies, http://www.readwriteweb.com/archives/semantic web patterns.php, 2008.

[17] ***, Muckety - Mapping relations and measuring influence, http://muckety.com/.

[18] ****, NNDB Mapper: Tracking the entire world, http://mapper.nndb.com/.