A Semantic Content Management System for e-gov applications

Donato Cappetta[1], Salvatore D'Elena[1] , Vincenzo Moscato[2], Vincenzo Orabona[1], Raffaele Palmieri[1], Antonio Picariello[2]

[1]*Eustema Spa, Via Carlo Mirabello, 7, 00195, Roma, Italy*

[2] *University of Naples Federico II, DIETI, via Claudio 21, 80125, Napoli, Italy*

{*d.cappetta,s.delena,v.orabona,r.palmieri*}*@eustema.it,*{*vmoscato,picus*}*@unina.it*

Keywords: CMS, Semantic Web, Ontologies, LOD

Abstract: In this paper, we describe a novel *Semantic Content Management System* (SCMS) able to handle multimedia contents of different kinds (e.g. texts and images) using the related semantics and capable of supporting e-gov applications in different scenarios. All the information is described using semantic metadata semi-automatically extracted from multimedia data, which enriches the browsing experience and enables semantic contents' authoring and queries. To this aim, several Semantic Web technologies have been exploited : RDF/OWL for data modeling and representation, SPARQL as querying language, Multimedia Information Extraction techniques for content annotation, W3C standard models, vocabularies and micro-formats for resource description. In addition, we propose for entity annotation issue the LOD approach. As an application scenario of the platform, we report a system customization useful for managing the semantic matching between the required professional profiles by a Public Administration and the available skills in a set of curricula vitae with respect to a given call.

1 Introduction

In spite of the widespread diffusion and use in a large variety of applications of CMS (Boye, 2012), nowadays the existing tools still lack consistent and scalable annotation mechanisms that allow them to deal with semantics of the managed contents with respect to heterogeneous application scenarios, especially concerning e-government applications.

As for the Web, the last generation of CMS focuses their attention on data (information embedded in a document) rather than content (the document itself), thus shifting from a "content centric" vision to a "data centric" one.

The data centric approach is then endorsed by *Enterprise Information Management* (Van Til et al., 2010) and *Linked Data* (Linked Data, 2011) paradigms, which state as data and associated meaning can independently live respect to the applications, allowing their interoperability in according to the *Semantic Web* issues.

Recently, in according with this new trend some CMS and wiki systems, such as Drupal RDF module or the RDF Tools for Wordpress (García et al, 2008), have started to incorporate semantic annotation modules in order to cope with the described lack.

However, all these initiatives do not yet provide a fully featured semantic CMS, especially if one considers the different kinds of content beyond the HTML documents.

Generally, in the CMS context, the introduction of a semantic model able to represent and manage contents' semantics can be supported by the development of reusable software components assembled within a *Semantic Framework* (SF), useful to build different vertical applications in several domains (see Fig. 1).

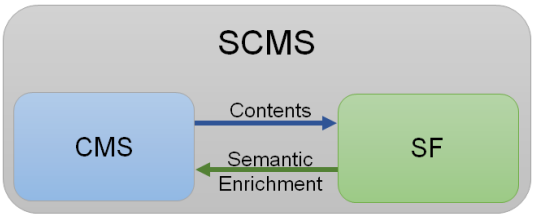

Figure 1: CMS and SF.

In this paper, we present an on-going research project leaded by University of Naples and Eustema Company for the design and development of a novel *Semantic Content Management System* (SCMS), within a FIT call recently founded by the Italian Technology Innovation Ministry.

In particular, the project aims at realizing of a novel CMS capable of improving user experience in managing contents in several domains by means of a set of semantic facilities. We provide a CMS combined with a fully featured semantic metadata repository with reasoning capabilities. Both components, the CMS and the semantic repository, are integrated in a transparent way for the end-users and enable more sophisticated and usable interactions.

The paper is organized as in the following. First, we introduce the Content Lifecycle model of the proposed solution, second we detail the reference architecture for the implementation stage with implementation details and, finally, we present in a real life scenario an application of our SCMS for the e-gov domain.

2 Content Lifecycle Model

The proposed solution adopts for managed contents the lifecycle model depicted in Fig.2.

The model allows to describe information extracted from contents' in the RDF format and in according to the *Web of Data* Best Practices and Issues defined by W3C Consortium (Web of Data, 2014) .

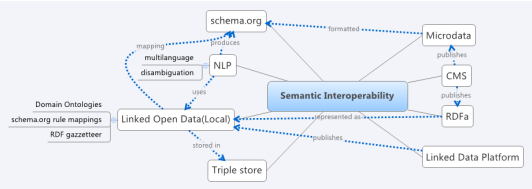

Figure 2: Content Lifecycle Model.

In a preliminary stage, we are able to extract several information from textual contents in the shape of Schema.org "tags" (Schema.org, 2011) through the application of a particular *Natural Language Processing* (NLP) pipeline (Bandyopadhyay et al., 2013), thus supporting a sort of *Entity Annotation and Linking* process. This step is very important, because it allows to infer and create links between terms extracted from the contents and their related meanings (e.g., "Paris" can be linked to "http://dbpedia.org/page/Paris"), using available public *Linked Open Data* (LOD) (Heath, 2013) information.

Furthermore, Schema.org tags are embedded into HTML fragments to increase *Google* or *Yahoo* search engines' performances in retrieving the related web pages. To this aim, *HTML Microdata* (HTML Microdata, 2013) and *RDFa* (RDFa, 2013) technologies have been exploited.

Frm the other hand, domain ontologies are opportunely used to map more specific and application-dependent terms with the related domain concepts by means of *RDFS Schema.org* vocabulary (Schema.org, 2011), representing entities and their relationships within of the ontology instance. In addition to public LOD entities, we inherit from W3C Consortium other ontological schema models as the *Ontology for Media Resources* one, used to represent metadata of the correlated multimedia description such as images or videos.

The final and obtained knowledge represented by a set of triples is finally stored in a Triple Store System and a reasoning layer is built on the top of it to produce new knowledge by using inference rules. An internal search engine has been developed to index extracted data and their URI, supporting search activities performed by users.

3 System Overview

3.1 Main Goals

The added value of the proposed semantic CMS lies in the capability of associating each managed content with a set of additional information which allow to derive its semantics and with the application domain by exploiting the *linked entities*.

In particular, entities are used to create relations among managed documents in CMS, and if they have references to LOD ontologies, the relations could be extended to all public documents on the web which deal with a similar topic.

Extracted entities are also used in the topic categorization process of contents - useful for automatic document classification aims - that uses a vocabulary of terms, already available for a given thematic domain and coded in the shape of taxonomies or thesauri.

3.2 Reference Architecture and System Functionalities

We decided to adopt for our system the reference architectural model reported in Fig.3.

From a functional point of view, the proposed system is partially inspired to the *Apache Stanbol* one (Apache Stanbol, 2013) and is based on a multilayer architectural pattern .

The basic provided system functionalities are: (i) *Administration and Configuration*, (ii) *Content Editing & Semantic Lifting* and (iii) *Semantic Search*.

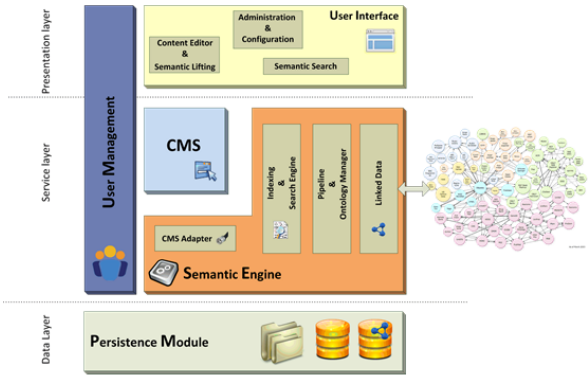

Figure 3: System Architecture.

The Administration and Configuration functionalities allow to:

- manage the available domain ontologies, taxonomies and vocabularies, related to the considered application domain.
- implement a set of rules to produce by proper reasoning mechanisms new derived and useful knowledge;
- associate LOD to domain entities.

If any knowledge source is available for the considered domain, users can eventually create a new ontology, a specific taxonomy, a custom vocabulary, etc. or extend some of existing ones and add them to the system Knowledge Base.

This step is performed in an off-line manner using some external tools that facilitate the production of all these kinds of resources (e.g. Protegé[1], Thesaurus Manager [2], etc.).

Content Editing and Semantic Lifting functionalities allow during contents' editing process to:

- link the typed text with existing entities in the knowledge base, or suggest some new entities;
- classify the topic of content with respect to a reference taxonomy;
- map each identified entity with the related LOD;
- validate in an interactive way entities and their relations, semantically extracted from the contents, before saving content with semantic enrichments;
- include in the web contents' publishing step the semantic annotation in terms of microformats and RDFa within the produced HTML;
- obtain the entity linking and topic classification of metadata related to multimedia contents.

[1]http://protege.stanford.edu/

[2]http://thmanager.sourceforge.net/

The Content Editing process and Semantic Lifting have been realized using RDFaCE [3] with its TinyMce Editor[4].

The choice of the first tool has been driven by the availability of some content annotation functionalities using RDFa and microdata. From the other hand, TinyMce Editor represents a valid choice because its a well known web based Javascript WYSIWYG editor, platform independent and extensively used within many open source CMS; furthermore it provides a clear set of API to extend its features with custom behaviors.

Semantic Search capabilities allow to:

- implement full text and faceted search;
- implement a semantically enriched search using concepts which are expandable according to predetermined relations (e.g. search the products through the company that produces them);
- implements the search of multimedia data similar to a given content;
- view and search the contents starting from LODs;
- browse contents and facts present in the knowledge base using SPARQL endpoint.

3.3 Implementation Details

The presentation layer has been implemented as a stand-alone client-side component, that communicates with the RESTful service layer via Ajax.

This component could be integrated in different CMS in a very easy way: for Liferay CMS, for example, the integration has been realized producing a customized Portlet.

The Persistence Layer implements storage and retrieval functionalities and manages two different kinds of information:

- CMS data ,
- SF data that are represented by the ontologies, vocabularies and all the resources used by the process of content semantic enrichment.

Semantic information are handled by a Triple Store System.

The technological choice, in this case, has fallen in the mixed use of Apache Clerezza [5] with OpenLink Virtuoso.

OpenLink Virtuoso presents an hybrid architecture that provides a set of capabilities, covering the following areas:

[3]http://rdface.aksw.org/

[4]http://www.tinymce.com/

[5]http://Clerezza.apache.org

- Relational Data Management,
- RDF and XML Data Management,
- Free Text Content Management and Full Text Indexing,
- Document Web Server,
- Linked Data Server,
- Web Application Server,
- Web Services Deployment.

The module that links the CMS and SF is the *CMS Adapter*, through which the managed content is synchronized the with extracted information during entity linking processing phase, together with other metadata.

There are also different access rights to the content that are necessary to guarantee the confidentiality of documents and are used by semantic search to filter results depending on the user performing the search.

The Development activity has moved across several directions. First, we have based the implementation of the discussed content model on *Apache Stanbol* components.

NLP tools have been then integrated to deal with contents in Italian language, not natively supported by Stanbol. In particular, we have used *Freeling* (Freeling, 2013) an open source suite of some language analyzers.

Another considerable development effort has regarded the CMS Adapter component for the integration of Apache Stanbol with CMS and the storage of semantic metadata.

We have used *Liferay* (Liferay) to transfer the contents to the *Apache Stanbol Content Hub* component in the *CMIS AtomPub* format, using a *RESTful* approach (see Fig.4).

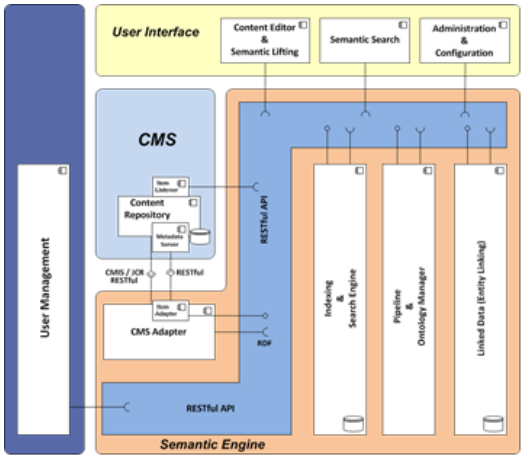

Figure 4: CMS Adapter.

A second customization has interested *Apache SOLR* (Apache Solr, 2010) component, the internal indexing and search engine that is able to manage metadata embedded into content, as well as the extracted text. To this aim, we have modified the NLP chain to add RDF formatted metadata within the pipeline output.

A further extension has regarded the *Emir* (Lux, 2009) integration, an open source tool for image annotation and similarity search. It represents image metadata in *MPEG7* format and translate them into Ontology for Media Resource entities.

In a nutshell, our SCMS implementation is WEM oriented. In fact, semantic lifting allows to integrate in a Content Editor GUI all the functionalities to suggest appropriate contents to the user, depending on what he is looking for at that time. Moreover, we increase search user experience thanks to the possibility of querying also semantic metadata, together with entity-based faceted search. By means of semantic query expansion mechanisms, it is possible to add related keywords for query execution and to produce more accurate results.

These keywords depend on then managed knowledge base, and on the available domain ontologies (for example, a query search for a particular disease was expanded by using a word of a drug for its treatment). Another kind of functionality of Semantic Engine is the Document Classification, where topics are listed in a proper taxonomy. In the case of CMS is able to profile user, depending on its search history, or more visited pages or feedbacks, classification could be used for suggesting to user the most relevant information for his/her preferences.

4 A Real-Life Scenario

About the possible applications of SCMS, there are several alternatives.

In the Enterprise context, for example, we could apply the interoperability model to many of legacy systems of the IT infrastructure as CRM, HR, ERP and so on. The ontology model built for representing all these data should be unique, thus we could design new business processes which can merge all information together and create a single point of view (*LED*, *Linked Enterprise Data* (Lacorix, 2013)).

In Big Data Analytics field, exploring newspaper articles to extract entities, facts and relationships, it should be possible to assess clients or suppliers reputation; by the analysis of social interactions, to anticipate clients expectations; by the insurance policies analysis, to prevent frauds via predictive algorithms.

As real-life application scenario of our SCMS platform, we report a system customization useful for managing the semantic matching between the required professional profiles by a Public Administration (PA) and the available skills in a set of curricula vitae with respect to a given call in the ICT area.

More in details, the PA employees need to verify the correct matching between the professional profiles and the skills reported in the curricula that participants have submitted for a public tender, with respect to the required profiles: this facility has to help the scoring process of competitors for the tender.

The first step consists of the system knowledge base building that has to represent and model the typical skills and professional profiles in the ICT context.

To this aim, we created the knowledge base starting from the development of a thesaurus of professional profiles - we use the *EUCIP* (EUCIP) classification - then enriched with the skills reported within the *DISCO II* (DISCO II) available thesaurus.

EUCIP (European Certification of Informatics Professionals) is the European standard for describing skills of ICT professionals.

DISCO, the European Dictionary of Skills and Competences, is an online thesaurus that currently covers more than 104,000 skills and competence terms and approximately 36,000 example phrases. Available in eleven European languages, DISCO is one of the largest collections of its kind in the education and labour market.

The DISCO Thesaurus offers a multilingual and peer-reviewed terminology for the classification, description and translation of skills and competences. It is compatible with European tools such as Europass, ESCO, EQF, and ECVET, and supports the international comparability of skills and competences in applications such as personal CVs and e-portfolios, job advertisements and matching, and qualification and learning outcome descriptions.

The construction of the knowledge base has been realized by defining a new ontology and a new thesaurus that considers the EUCIP ICT professional profiles and enriches them with the skills present in the DISCO II thesaurus, defining at the same time proper relations among such entities.

For this purpose, we have been supported by a domain expert in order to establish the right relationships between skills and profiles, and to validate them. In the following, we describe the necessary steps to accomplish the annotation process of resumes.

1. Resumes submitted by contractors are loaded into the SCMS platform through the User Interface, and in particular, exploiting the described Content Editing and Semantic Lifting facilities.
2. The Semantic Engine semantically enriches each received content: it analyzes the text and, through the execution of the NLP pipeline, provides the Entity Annotation process. Through the Linking process, extracted entities are then linked to well-known entities of the reference domain (that in this scenario are properly represented by professional skills). The obtained semantic information is finally then stored, together with the related resume, and indexed for the Semantic Search purposes.
3. The User Interface shows the results of the Semantic Lifting obtained through the Annotation process application, highlighting the words that cover a certain skill and showing the related professional profiles.

In order to provide a set of facilities for resumes' validation, the SCMS has been equipped with a functionality that allow users to check if the skills and the professional profiles match with those ones required by the tender.

The User Interface (see Fig. 5) shows how the user can easily retrieve the correspondence between the skills resumes and the professional profiles. In particular, in the same view, it is possible to show the required skills together with those ones present in the resume, but not necessarily desired.

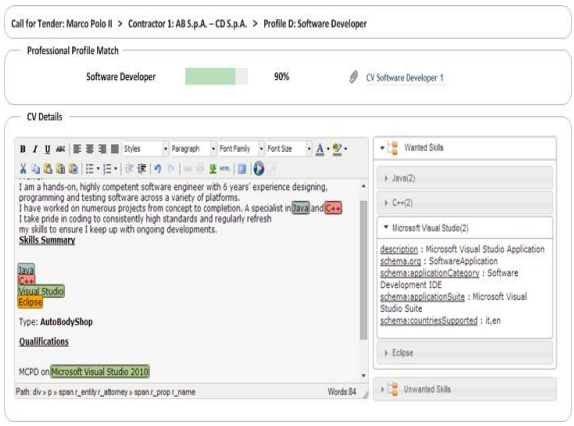

Figure 5: Resume Analysis.

The professional profile and skills - that the Semantic Engine has inferred - are then compared with the required ones showing the percentage amount of matching, calculated as a confidence parameter (see Fig. 6).

This simple business scenario, regarding e-government applications, can also be applied to other cases, concerning the composition of a work team at the start of new incoming projects in an ICT company, for example.

Following the definition of certain profiles, required for developing the project, users can search all the resumes in a corporate database, to determine which ones match with the specific requirements.

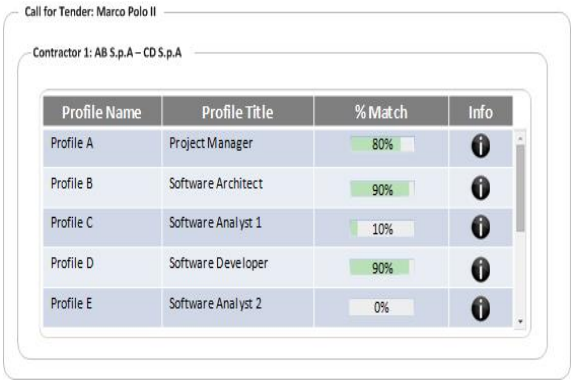

Figure 6: Semantic Matching Results.

REFERENCES

Apache, Apache Solr Online Available: https://lucene.apache.org/solr/, 2010.

Apache, Apache Stanbol. Online Available: https://stanbol.apache.org, 2013.

S. Bandyopadhyay et al. Emerging Applications of Natural Language Processing: Concepts and New Research. Information Science Reference, 2013.

J. Boye, What's in a name. Online Available: http://www.slideshare.net/JanusBoye/whats-in-a-name-what-do-we-really-mean-with-cms-in-2012., 2012.

U. P. d. Catalunya, Freeling. Online Available: http://nlp.lsi.upc.edu/freeling/, 2013.

FBK, Web of Data. Online Available: http://wed.fbk.eu/, 2014

T. Heath, LinkedData.org. Online Available: http://linkeddata.org/, 2013.

Lacroix, Linked Enterprise Data Online Available: http://www.inria.fr/content/.../Fabrice-LACROIX.pdf, 2013.

Liferay. http://www.liferay.com/.

M. Lux, Semantic Metadata. Online Available: http://www.semanticmetadata.net/features/, 2009.

W3C, Linked Data. Online Available: http://www.w3.org/egov/wiki/Linked_Data, 2011.

W3C, Schema.org. Online Available: http://www.schema.org, 2011

W3C, HTML Microdata. Online Available:http://www.w3.org/TR/microdata/, 2013.

W3C, RDFa. Online Available: http://www.w3.org/TR/xhtml-rdfa-prime, 2013.

W3C, SPARQL. Online Available: http://www.w3.org/TR/sparql11-overview/, 2013.

P. Van Til, A. van der Lans, P.l Baan, Enterprise Information Management.

Book, Lulu.com, 2010.

Roberto García, Juan Manuel Gimeno, Juan Manuel, Ferran Perdrix, Rosa Gil, Marta Oliva, The rhizomer semantic content management system, Emerging Technologies and Information Systems for the Knowledge Society, pp. 385–394, 2008, Springer

EUCIP, EUCIP Profiles. Online Available: http://www.eucip.it/.

3s Unternehmensberatung (AT), DISCO 2 Project. Online Available: http://disco-tools.eu/disco2_portal/