# CMS towards Semantic Interoperability

Donato Cappetta[1], Vincenzo Moscato[2], Vincenzo Orabona[1], Raffaele Palmieri[1], and Antonio Picariello[2]

[1] Eustema Spa, Via Carlo Mirabello, 7, 00195, Roma, Italy
{d.cappetta,v.orabona,r.palmieri}@eustema.it,
[2] University of Naples Federico II, DIETI, via Claudio 21, 80125, Napoli, Italy
{vmoscato,picus}@unina.it

**Abstract.** In this paper we present our development experience of a Semantic Content Management System able to handle and manage uniformly heterogeneous contents of different kinds (texts, video and images) considering the related semantics. To this aim, we exploit several Semantic Web technologies: RDF/OWL for data modeling and representation, SPARQL as querying language, Multimedia Information Extraction techniques, W3C standard models for creating taxonomies and resource annotation, vocabularies and microformats. We also follow the Best Practices and Issues for the Web of Data by reusing LOD information for the Entity annotation of the managed content, thus providing CMS with advanced capabilities such as contents' authoring and tagging.

## 1 Introduction

This paper discusses 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 an interesting keynote [4], the meaning of term *CMS* has been discussed, showing the related recent evolutions for such kind of systems. Some acronyms correlated to the concept of CMS, such as *WCM* (*Web Content Management*), *ECM* (*Enterprise Content Management*), *WEM* (*Web Experience Management*) have been analyzed and contextualized with respect to several features and general issues:

- Portals, Responsive, Search, Social Collaboration ,
- Digital Asset Management, Multimedia Asset Management, Digital Marketing,
- Content Integration, Analytics and Big Data,
- Blogging, Digital Workplace, Mobile, Gamification.

This variety of functionalities proves the large amount of potential applications of a CMS that furthermore depend on the type of managed contents, the granularity of data elaboration process and the supported collaboration level among users for the creation of new content.

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* [16] and *Linked Data* [11] 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.

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

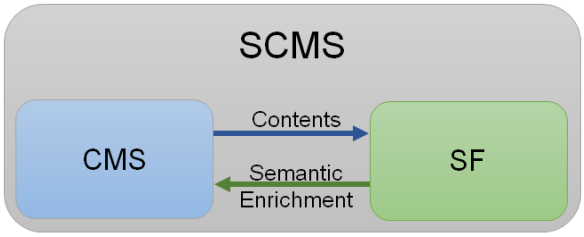


**Fig. 1.** CMS and SF.

In this paper, we describe the SCMS project concerning the realization of a CMS capable of improving user experience by means of a set of semantic facilities. First, we introduce the interoperability model of the proposed solution, second we detail the reference architecture for the implementation stage and, finally, we present the new features derived by the use of semantic technologies together with some implementation details and possible applications.

## 2 Semantic Interoperability Model

The proposed solution adopts the semantic interoperability model depicted in Fig.2. The model allows to describe information extracted from contents' in the *RDF* format and in according to the *W3C* issues.

In a preliminary stage, we are able to extract information from textual contents in the shape of *Schema.org* "tags" [12] through the application of a *NLP* (*Natural Language Processing*) pipeline [3], 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").

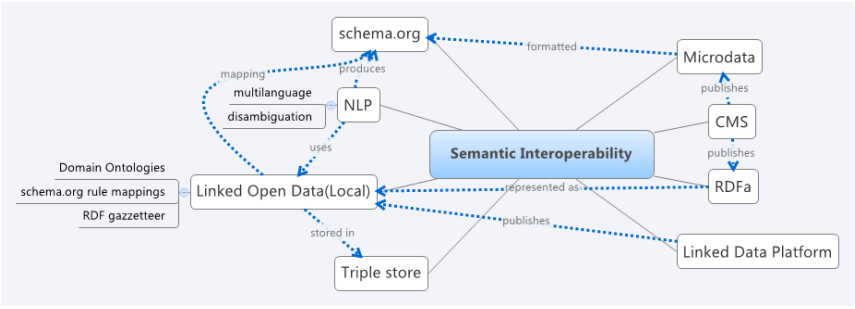


**Fig. 2.** Semantic Interoperability Model.

To perform this activity, we follow the best practices and issues for *Web of Data* [6] that suggest to reuse available public *LOD* (*Linked Open Data*) [7] information for the entity annotation. 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* [13] and *RDFa* [14] technologies have been exploited.

Domain ontologies are then opportunely used to map more specific and application-dependent terms with the related domain concepts, by means of *RDFS Schema.org* vocabulary [12] to represent 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. Eventually, we expose linked data generated by the CMS by means of a *SPARQL* endpoint [15], which could be used from third party applications for querying the system and browsing our data. An internal search engine has been developed to index extracted data and their *URI*, supporting search activities performed by users.

## 3 The Reference Architecture

During design activities, the proposed interoperability model has been traduced into the reference architectural model reported in Fig.3.

The architecture (in terms of some components of logical view and software artifacts) is partially inspired to the *Apache Stanbol* [2].

In this paper, we present only a part of the implemented functionalities. In particular, we focus our attention on the *User Interface* module that provides the interfaces and facilities through which a user can access to all the system functionalities: contents' editing, semantic analysis and search and configuration.

More in details, the described module is then composed by the following components:

- *Content editor and Semantic Lifting*:
  - in the contents' editing process allows to link the significant terms (from text or metadata) with entities of an existing vocabulary;
  - in the web contents' publishing step allows to inserts the semantic annotaion in terms of microformats and RDFa within the produced HTML;
  - allows to show classification of contents in terms of topics and with respect to a reference taxonomy;
  - allows to map each identified entity with the related LOD;
  - allows to validate entities and their relations, semantically extracted from the contents;
  - allows to how and classify all the multimedia description related to a content.
- *Administration and Configuration*
  - allows to manages available domain ontologies and taxonomies;
  - allows to implement a set of reasoning rules to produce new derived and useful knowledge;
  - associates LOD to domain entities
- *Semantic search*
  - implements full text and faceted search;
  - implements 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 documents and multimedia content similar to a given content;
  - allows to view and search the contents starting from LODs;
  - allows to search contents browsing graph of concepts (SPARQL endpoint).

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.

## 4 Implementation Details and Possible Applications

The added value of the proposed semantic CMS lies in the capability of associating each managed content with a set of additional information which consent to derive its semantics and 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.

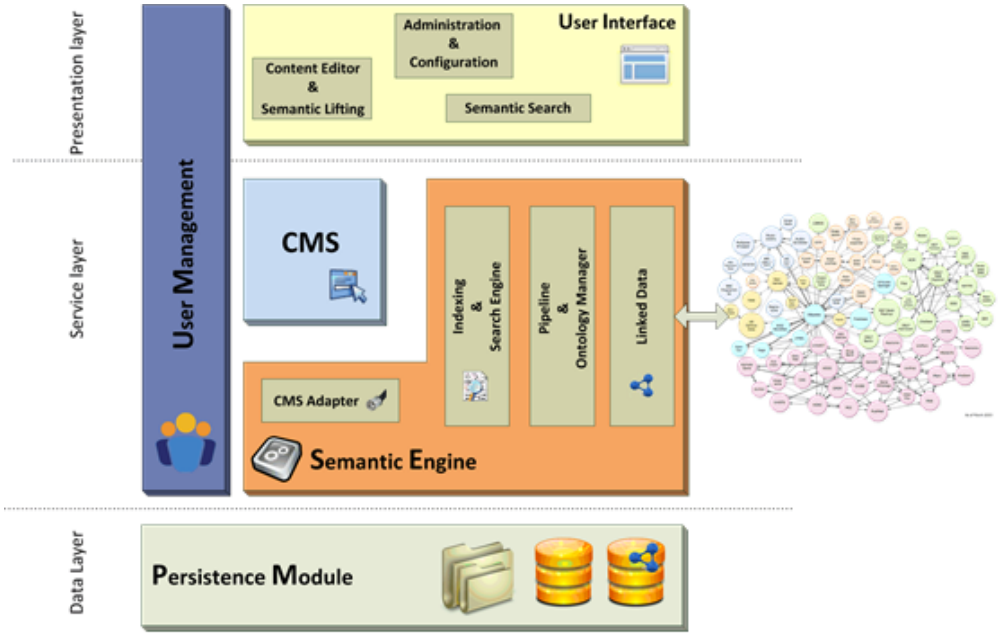


**Fig. 3.** System Architecture.

Extracted entities are also used in the topic categorization process of contents that uses a vocabulary of terms, already available for a given thematic domain and coded in a taxonomy or thesaurus shape.

The Development activity has moved across several directions. First, we have based the implementation of the discussed interoperability 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* [5] 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* [9] to transfer the contents to the *Apache Stanbol Content Hub* component in the *CMIS AtomPub* format, using a *RESTful* approach (Fig.4).

A second customization has interested *Apache SOLR* [1] component, the Stanbol 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 into pipeline output. A further extension has regarded the *Emir* [10] 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.

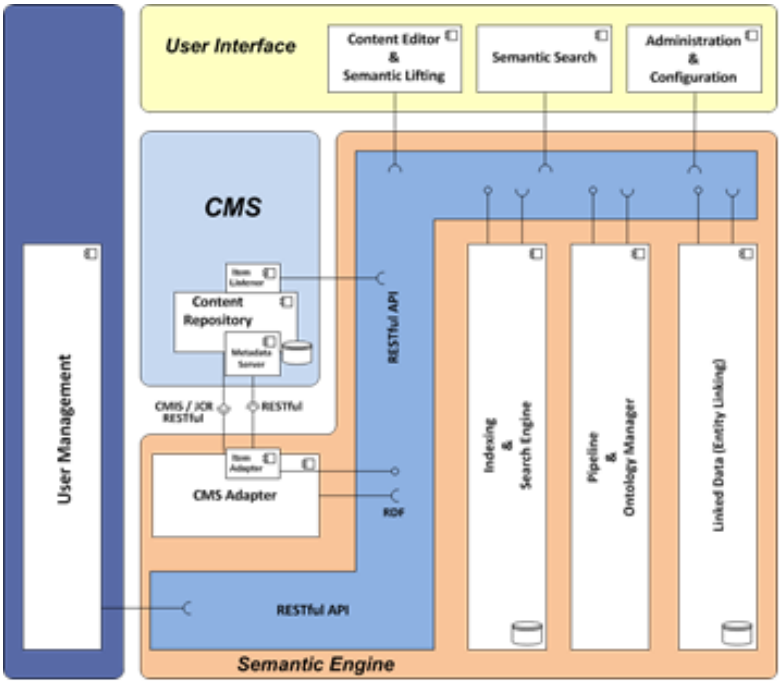


**Fig. 4.** CMS Adapter.

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.

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* [8]).

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.

## References


1. Apache, Apache Solr Online Available: https://lucene.apache.org/solr/, 2010.
2. Apache, Apache Stanble. Online Available: https://stanbol.apache.org, 2013.
3. S. Bandyopadhyay et al. Emerging Applications of Natural Language Processing: Concepts and New Research. Information Science Reference, 2013.
4. 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.
5. U. P. d. Catalunya, Freeling. Online Available: http://nlp.lsi.upc.edu/freeling/, 2013.
6. FBK, Web of Data. Online Available: http://wed.fbk.eu/, 2014
7. T. Heath, LinkedData.org. Online Available: http://linkeddata.org/, 2013.
8. Lacroix, Linked Enterprise Data Online Available: http://www.inria.fr/content/download/18171/514403/version/1/file/Antidot_Fabrice-LACROIX.pdf, 2013.
9. Liferay. http://www.liferay.com/.
10. M. Lux, Semantic Metadata. Online Available: http://www.semanticmetadata.net/features/, 2009.
11. W3C, Linked Data. Online Available: http://www.w3.org/egov/wiki/Linked_Data, 2011.
12. W3C, Schema.org. Online Available: http://www.schema.org, 2011
13. W3C, HTML Microdata. Online Available:http://www.w3.org/TR/microdata/, 2013.
14. W3C, RDFa. Online Available: http://www.w3.org/TR/xhtml-rdfa-prime, 2013.
15. W3C, SPARQL. Online Available: http://www.w3.org/TR/sparql11-overview/, 2013.
16. P. Van Til, A. van der Lans, P.l Baan, Enterprise Information Management. Book, Lulu.com, 2010.