An Application of Semantic Web Technologies to Enhance Content Management in Web Information Portals

Vincenzo Orabona[1], Raffaele Palmieri[1], Vincenzo Moscato[2], Antonio Picariello[2], Salvatore D'Elena[1] and Donato Cappetta[1]
[1]*Eustema Spa, via Carlo Mirabello 7, 00195, Roma, Italy*
[2]*University of Naples, Federico II, DIETI, via Claudio 21, 80125 Napoli, Italy*

Keywords: CMS, Semantic Web, LOD.

Abstract: As well known, *Semantic Web* technologies make available a set of facilities that allow data to be shared and reused across applications. The last generation of *Content Management System* (CMS) can leverage such technologies to improve the content management task incorporating semantic annotations of the produced resources. Here, we present the benefits deriving from the application of semantic technologies in a CMS environment. To this goal, we collect preliminary results about the effectiveness of the integration of a semantic annotation engine within the *Intrage Web Portal* for content management purposes. The obtained results show that the approach is quite promising and encourage the current research.

1 INTRODUCTION

As well known, *Semantic Web* technologies make available a set of facilities for enabling interoperability among software agents in the Web, providing a common framework that allows data to be shared and reused across applications. From the other hand, the related data formats (as XML and RDF) constitute a suitable mean to represent in a machine understandable way the knowledge connected to the great amount of semi-structured or unstructured documents accessible by the Web itself.

Following the Semantic Web vision, the last generation of *Content Management System* (CMS) focuses their attention on data (information embedded in a document) rather than content (the document itself), thus shifting from a "content centric" approach to a "data centric" one (Alalwan and Weistroffer, 2012; Garcìa et al., 2008).

To this goal, they incorporate semantic annotation modules in order to derive useful information from the managed contents and deal with their semantics, leveraging the *Linked Data* paradigm to relate extracted concepts with the available external knowledge (often coded in the shape of vocabularies, taxonomies or ontologies) depending on the considered application scenario (Dalkir, 2013).

Indeed, taxonomies, vocabularies of terms as well as ontologies with their relationships, also represent a valuable source for supporting the *Natural Language Processing* (NLP) tools to extract information, depending on the specific domain (Bandyopadhyay, 2012, De Virgilio et al., 2012).

In a recent paper (Cappetta et al., 2014), we described the design and development of a novel *Semantic Content Management System* (SCMS), representing our solution to the content management processing problem. In particular, we provided a CMS combined with a fully featured *semantic metadata repository* with reasoning capabilities.

In a preliminary stage, we are able to extract several information from textual contents in the shape of *Schema.org* "tags" through the application of a particular NLP pipeline, thus supporting a sort of *Entity Annotation and Linking* process (Sagayam et al., 2012).

Furthermore, the tags are embedded into HTML fragments to increase search engine performances. In addition, domain and *Linked Open Data* (LOD) ontologies (e.g. DBpedia and Geonames) are then used to map specific terms with the related domain or concepts by means of RDFS vocabularies. Eventually, *Ontology for Media Resources* was exploited to represent metadata of the correlated multimedia description such as images or videos (Amato et al., 2009).

Orabona V., Palmieri R., Moscato V., Picariello A., D'Elena S. and Cappetta D..
An Application of Semantic Web Technologies to Enhance Content Management in Web Information Portals.
DOI: 10.5220/0005579803780381
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (KomIS-2015), pages 378-381
ISBN: 978-989-758-103-8

The final and obtained knowledge represented by a set of triples in the RDF format is finally stored in a *Triple Store System* and a reasoning layer was built on the top of it to produce new knowledge by using proper inference rules (coded in the SWRL format).

An internal search engine has been developed to index extracted data and their URI, supporting semantic retrieval by means of query expansion and query by example (based on a notion of *Document Similarity*) mechanisms. Another functionality of our semantic engine is the *Document Classification*, where related topics are listed in a proper taxonomy.

Summarizing, in the following list we report the features offered by the developed system:

1. **Entity Extraction:** extraction of entities belonging to particular classes from a text;
2. **Conceptualization:** extracting semantic concepts from texts, summarizing the analyzed content;
3. **Classification:** classification of structured (for example contained in a database) and unstructured information with respect to a reference model, such as a taxonomy, supported by inference mechanisms, also, and not purely *keyword-based*;
4. **Relations Extraction:** ability in automatically relations between concepts reported in a given text**;**
5. **Language Detection**: capability of identifying the language used in a text;
6. **Document Similarity:** given a document (e.g. an image) or a portion of text, the capability of recommending similar resources to the target one based on the analysis of their contents;
7. **Query expansion:** expanding the initial keyword used to search documents, with a set of "related terms" (extracted from the available ontologies);
8. **Linked Open Data:** linking the extracted entities to public datasets to facilitate the search of documents from the Web and to improve ranking in search engines**.**

The reference architecture with all the implementation details are reported in (Cappetta et al., 2014).

In this work, we want to report an application of the described SCMS to a real case study: the *Intrage Portal*. In particular, our aim is to show the advantages related to the application of semantic technologies for managing contents within a Web Information Portal.

2 THE INTRAGE PORTAL

The Intrage Web Portal (http://www.intrage.it/) was created to help "over 50" people to fulfill all those needs concerning with their life and to have some suggestions about future issues such as: retirement and social assistance.

More in details, the topics provided with the Intrage Portal are: jobs, retirements, health, insurance, taxation, social welfare, homelife, consumer (Figure 1 shows the related home page).

On line since 16 June 2000, we have published more than 100,000 contents among pages of news, analytical services, special articles and interviews. During these years, the web site has received about 100 million of visits and 300 million of page views. Moreover, it annually produces 4,000,000 newsletter and more than 12,000 lonely hearts have published

their cards in the section "Soul Mate".

Figure 1: Home Page of the Intage Portal.

Intrage experience was reported by about 100 newspapers (national daily newspapers, periodicals and on-line news sites) and by several television networks, with more than 280 assets between articles and interviews.

In the following, we briefly describe how we have used our SCMS to improve the management of the contents within the portal.

The first and fundamental step to enable the relevant information extraction and correct content classification functionalities is the definition of a domain ontology for over 50 people interests.

In the specific case of Intrage, the ontology has been represented by a *SKOS taxonomy*, given the large variety of the available concepts. The elements useful to populate the taxonomy, were collected starting from the thematic channels already provided by the web portal. Such elements constitute the "top" concepts, and were enriched with other concepts provided by the *Nuovo Soggettario taxonomy*, published by the Library of Florence (http://thes.bncf.firenze.sbn.it/).

Eventually, particular SKOS propertieshave been used properties to link the different concepts, together with some references to *Wikipedia*.

Successively, the discussed entity annotation and linking process was activated to create the final System *Knowledge Base* in according to the LOD paradigm (Amato et al., 2012). Semantic retrieval facilities are finally available to find contents of interest both for editing and browsing/searching aims. In the Figure 2, we report a fragment of the adopted SKOS taxonomy.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#" >
  <rdf:Description rdf:about="http://na-modern:8890/scms/intrage/licenziamento">
    <skos:narrower rdf:resource="http://na-modern:8890/scms/intrage/licenziamento-ingiustificato"/>
    <skos:narrower rdf:resource="http://na-modern:8890/scms/intrage/statuto-dei-lavoratori"/>
    <skos:prefLabel xml:lang="it">Licenziamento</skos:prefLabel>
    <skos:broader rdf:resource="http://na-modern:8890/scms/intrage/rapporto_lavoro"/>
    <skos:narrower rdf:resource="http://na-modern:8890/scms/intrage/preavviso-di-licenziamento"/>
    <skos:related rdf:resource="http://na-modern:8890/scms/intrage/cigs"/>
    <skos:notation rdf:datatype="http://dewey.info">331.2596</skos:notation>
    <skos:related rdf:resource="http://na-modern:8890/scms/intrage/cigo"/>
    <skos:related rdf:resource="http://na-modern:8890/scms/intrage/indennità-di-disoccupazione"/>
    <skos:inScheme rdf:resource="http://iaaa.unizar.es/thesaurus/intrage"/>
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <skos:related rdf:resource="http://na-modern:8890/scms/intrage/lavoro-subordinato"/>
    <skos:related rdf:resource="http://na-modern:8890/scms/intrage/licenziamento-ad-nutum"/>
  </rdf:Description>
```

Figure 2: SKOS taxonomy.

3 OBTAINED ADVANTAGES

In the following, we summarize all the main advantages derived by the application of semantic technologies to the content management problem for the proposed case study.

1. **To Facilitate the Content Editing task through the suggestion of similar previously published content, during the production step** – When new contents are generates some facilities are provided to users suggesting previously published contents with "similar" keywords or topics. Thus, the editor has the possibility to link to other sources and verify the "originality" of what is being produced.
2. **To Improve Content Annotation through the suggestion of a set of metadata and tags automatically extracted from the content itself** – In addition to *Named Entity Recognition* (NER) utilities, our system provides on the base of a statistical analysis of the text a set of metadata and tags that can be useful to describe a content. This information is then exploited in the indexing stage for creating an *index of terms*.
3. **To improve the "visibility" of the published pages by injecting in the HTML code particular tags for search engines** – In the Content Editor tool, the annotation results obtained from the text are integrated as *microdata* within the HTML page, using additional span tags. These annotations are built by reusing the standard vocabulary Schema.org. Since search engines are able to process this additional information, and exploit them in the results they produce, they can apply higher rating criteria for pages containing this data, increasing their visibility;
4. **To Provide effective search mechanisms for content produced by both editors and final users, allowing to perform queries according to criteria used in the information extraction stage** – During the production of new contents, a tool is provided for users to determine correlated articles in the Knowledge Base. In the search phase, users can browse contents using the index of terms and exploit relations among ontology concepts for improving search results.
5. **To Increase Revenue derived by the advertising content through its contextualization in web pages** - For this purpose, we used a contextual taxonomy composed by two levels for the categorization of content, defined by the *IAB Association* (www.iab.it).
6. **To Enhance User Experience using content recommendation facilities, based on user profile (favourite pages, subscribed channels, etc.)** - All recommendation systems exploit user profiles to provide suggestions about contents related to particular topics, concepts or entities Amato et al., 2014; Moscato et al., 2013). For the Intrage Portal, content recommendation is implemented in "My Home" section, where user can view targeted recommendation boxes. These are determined by an ad-hoc algorithm, which assigns a particular score to contents of interest on the base of user feedbacks about followed channels, favourite tags, etc.

4 PRELIMINARY EVALUATION

We have to try to "measure" the introduced benefits for final users derived from the application of semantic technologies to a CMS environment. In particular, measurements have been performed to "quantify" the effective utility of the discussed advantages.

To this aim, for each benefit we have defined a particular indicator using a "5 five stars" rating. The following table summarizes the obtained results.

Table 1: Preliminary Results.

Best Practices		Expected Results		Obtained Results
S1B1	Semantic classification of content	RA1	Better time-efficiency of editing and publishing content	★★★★
S1B2	Semantic tagging			
S1B3	Web Experience Management	RA2	Enhance user experience	★★★
S1B4	Semantic geo-referentiation	RA3	Enhance user experience, enable geospatial content search.	★★★★
S1B5	Editorial support	RA4	Enhance user experience of editor, due to semantic features that allow him to focus on own creativity for creating original content, enriched by semantic information.	★★★★
S1B6	SEO	RA5	The use of schema.org enhance visibility of published content on web search engines through well-known vocabulary.	★★★
S1B7	Faceted Browsing	RA6	Enhance user experience for search functionalities, implementing an assisted navigation of content, specialized on domain.	★★★
S1B8	Search expanded semantically	RA7	Enhance user experience for search functionalities using natural language, returning a broader set of results, but always belonged to the domain.	★★★

In particular, the Table shows the introduced best practices related to the use of semantic technologies, the expected results and, finally, the measured score combing human dependent and objective criteria.

5 CONCLUSIONS AND FUTURE WORK

In this work, we described an application of Web Semantic technologies to the content management problem for Web Information Portal, reporting a real case study: the *Intrage Portal*.

We listed the possible benefits in the content production and search, trying to measure their range for the discussed case study.

Future work will be devoted to enhance user profiling using clustering and co-clustering techniques, and content recommendation exploiting hybrid strategies taking into account also the features of suggested items.

REFERENCES

Alalwan, J. A., and Weistroffer, H. R. (2012). Enterprise content management research: A comprehensive review. *Journal of Enterprise Information Management*, 25(5): 441–461.

Garcìa, R., Gimeno, J. M., Perdrix, F., Gil, R., and Oliva, M. (2008). The rhizomer semantic content management system. *Emerging Technologies and Information Systems for the Knowledge Society*, pages 385–394, Springer.

Amato, F., Mazzeo, A., Moscato, V., and Picariello, A. (2009). Semantic management of multimedia documents for e-government activity. In *Proceedings of IEEE International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009)*, pages 1193–1198.

Amato, F., Chianese, A., Moscato, V., Picariello, A., and Sperli, G. (2012). SNOPS: a smart environment for cultural heritage applications. In *Proceedings of the twelfth international workshop on Web Information and Data Management (WIDM)*, pages. 49–56, ACM.

Amato, F., Mazzeo, A., Moscato, V., and Picariello, A. (2014). Exploiting cloud technologies and context information for recommending touristic paths. *Studies in Computational Intelligence (Intelligent Distributed Computing VII)*, pages 281–287, Springer.

Bandyopadhyay, S. (2012). Emerging Applications of Natural Language Processing: Concepts and New Research, IGI Global publisher.

Cappetta, D., D'Elena, S., Moscato, V., Orabona, V., Palmieri R., and Picariello, A. (2014). A Semantic Content Management System for e-Gov Applications. In *Proceedings of 3rd International Conference on Data Management Technologies and Applications (DATA 2014)*, pages 440–445.

Dalkir, K. (2013). *Knowledge management in theory and practice*, Routledge.

De Virgilio, R., Orsi, G., Tanca, L., and Torlone, R. (2012). Nyaya: A system supporting the uniform management of large sets of semantic data. In *Proceedings of IEEE 28th International Conference Data Engineering (ICDE 2012)*, pages 1309–1312.

Jonquet, C., Musen, M.A., and Shah, N. (2008). A system for ontology-based annotation of biomedical data. In *International Workshop on Data Integration in the Life Sciences (DILS 2008)*, pages 144–152.

Moscato, V., Picariello, A., and Rinaldi, A. M. (2013). Towards a user based recommendation strategy for digital ecosystems. *Knowledge-Based Systems*, 37:165–175.

Sagayam, R., Srinivasan, S., and Roshni, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal Of Computational Engineering Research*, 2(5).