Full Length Research Paper

SEAN: Multi-ontology semantic annotation for highly accurate closed domains

Juan Miguel Gómez-Berbís, Ricardo Colomo-Palacios*, José Luis López-Cuadrado, Israel González-Carrasco and Ángel García-Crespo

Department of Computer Science, Universidad Carlos III de Madrid, Av. Universidad 30, Leganés, 28911, Madrid, Spain.

Semantic annotation has gained momentum with the emergence of the current user-generated content paradigm on the Web. The ever-growing quantity of collaborative data sources has resulted in the requirement for efficient approaches to create, integrate and retrieve information more efficiently in an environment where the users ask for accurate information. The main research challenge of the current work is using manual semantic annotation in a highly accurate closed domain, a conceptual domain with a minimal set of concepts where the benefits of adding semantics, search efficiency, optimization and the cost estimations are viable. This paper presents a semantic annotation approach for highly accurate closed domain based on multi-ontology annotation (domain and application ontologies).

Key words: Semantic annotation, semantic search, highly accurate closed domains, multi-ontology.

INTRODUCTION

The shift from Web 1.0 to Web 2.0 produced a change in how information is delivered and produced. In such scenario, the advantages of "Web 2.0" have increased the interest of companies as a way to obtain benefits from this environment in terms of communication with their customers (Ferreira, 2010). But, the amount of information available carries out problems of information overload. Not in vain, our social connectivity might have even increase in importance in the last years simply by the virtue of the information overload we are facing (Mika, 2005). Semantic web is seen as one of the solutions to this problem, and expectations are high for the Semantic Web, because information overload currently reduces the Web's usability (Euzenat, 2002).

Naeve (2005) states that the Semantic Web has initiated a paradigm shift from "knowledge push" to "knowledge pull", as a result of its advanced capacities for automatic information integration. Similarly, Fensel and Musen (2001) consider the Semantic Web as "a brain for humankind" and some authors have even extended this definition to a "human semantic web"

(Naeve, 2005; Vossen et al., 2007). However, in order to adequately exploit the capacities of the Semantic Web, it is necessary to carry out semantic annotation of its contents. Semantic annotation is considered a promising technology to add and manage the knowledge associated with a set of resources. Particularly, annotating specific domains with accuracy from an automatic or semi-automatic viewpoint has raised a challenge for the current state of the art of semantic technologies. With the advent of the new user-generated paradigm encompassed by the Web 2.0 phenomenon, information sources spread across the Web at an exponential rate and organizations have begun to make their business functionalities explicitly available to users.

In spite of this, these information domains are highly complex due to their distributed, easily extendible and chaotic nature, introducing a challenge in the community to accurately define such domains. This paper introduces the notion of highly accurate closed domains (HACD) as a set of domains with a minimal semantic model of concepts, that is, a domain which can be very accurately defined by a set of concepts and hence can be very easily annotated manually. The present focus is on HACD, since they encompass the domains the researchers are currently working on, such as software

*Corresponding author. E-mail: rcolomo@inf.uc3m.es.

development projects or particular software engineering methodologies, such as the European Software Agency (ESA).

This paper proposes SEAN, a global framework for multi-ontology semantic annotation. This framework is based on the manual semantic annotation of documents associated with entities, here entitled projects. The products of the projects have in common the fact that they can be annotated semantically. Finally, SEAN is not using specific properties of HACD, but simply using a minimal set of concepts that describe properly a concrete domain.

STATE OF THE ART

Semantic technologies have been pointed out as the future of Web (Benjamins et al., 2008) and a new way to support knowledge (Vossen et al., 2007; Fensel and Musen, 2001) in a wide range of domains (Lytras and García, 2008). Semantic technologies, based on ontologies (Fensel, 2002), provide a common framework that enables for data integration, sharing and reuse from multiple sources. Durguin and Sherif (2008) portrays the Semantic Web as the future web where computer software agents can carry out sophisticated tasks for users.

This approach facilitates the integration of data coming from a broader non-relational domain of data, which additionally might be distributed and outside enterprise boundaries and control (García, 2010). Taking this into account, according to Alani et al. (2008), Semantic Web applications are beginning to be pragmatic. Technology journalist, Markoff (2006) begun to call this new web applications as Web 3.0 and this tendency was latter followed by others (Lassila and Hendler, 2007; Hendler, 2008; Wang, 2008; Hendler, 2009). Web 3.0 can bring a new breed of spectacular applications compared to Web 2.0 with the same magnitude that separates Web 2.0 from Web 1.0 (Cardoso, 2007).

Semantic technologies have emerged as a new and highly promising context for knowledge and data engineering (Vossen et al., 2007). The term "Semantic Web" was coined by Berners-Lee et al. (2001), to describe the evolution from a document-based web towards a new paradigm that includes data and information for computers to manipulate. The essential difference between the classic Web and the Semantic Web is that structured data is exposed in a structured way (Gruber, 2008). The Semantic Web provides an alternative solution to represent the comprehensive meaning of integrated information and promises to lead to efficient data management by establishing a common understanding by means of ontologies (Shadbolt et al., 2006).

Certainly, ontologies (Fensel, 2002) are the technological cornerstones of the Semantic Web. The Semantic Web enables automated information access based on machine-processable semantics of data. Being machine-processable means that semantic search services can be make information available for providing precise and exhaustive information retrieval (Guha, 2003). Ontologies provide information systems with a semantically rich knowledge base for the interpretation of unstructured content (Mikroyannidis and Theodoulidis, 2010).

The term "ontology" can be defined as "a formal and explicit specification of a shared conceptualisation" (Studer et al., 1998). Ontologies provide a common vocabulary for a domain and define, with different levels of formality, the meaning of the terms and the relations between them. Knowledge in ontologies is mainly formalized using five kinds of components: classes, relations, functions, axioms and instances (Gruber, 1993). Classes in the ontology are usually organized into taxonomies. Sometimes the definition of ontologies has been diluted, in the sense that taxonomies are considered to be full ontologies (Studer et al., 1998).

The theory which supports the use of ontologies is a formal theory within which not only definitions but also a supporting framework of axioms is included (Smith, 2003). Ontologies were developed in the field of artificial intelligence to facilitate knowledge sharing and reuse (Fensel et al., 2001). Languages such as Resource Description Framework (RDF) and Ontology Web Language (OWL) have been developed; these languages allow for the description of web resources, and for the representation of knowledge that will enable applications to use resources more intelligently (Horrocks, 2008). These languages, and the tools developed to support them, have rapidly become de facto standards for ontology development and deployment; they are increasingly used, not only in research laboratories, but in large scale IT projects (Horrocks, 2008).

The Semantic Web consists of several hierarchical layers, where the ontology layer, in form of the OWL Web Ontology Language (recommended by the W3C), is currently the highest layer of sufficient maturity (Lukasiewicz and Straccia, 2008).

Ding (2010) asseverates that Semantic Web is fast moving in a multidisciplinary way. Taking full advantage of ontologies, the Semantic Web provides a complementary vision as a knowledge management environment (Warren, 2006) that, in many cases has expanded and replaced previous knowledge and information management archetypes (Davies et al., 2007). The goals of the Semantic Web initiative include the integration of data from different sources in a machine-processable format in order to make them accessible to computer programs and facilitating the use of data in ways that have not been thought of when the data was entered or recorded (Battré, 2008). It is agreed that semantic enrichment of resources would lead to better search results (Scheir et al., 2008). Due to the

impact of semantic technologies, there are many service areas in which these semantic technologies are being adopted: Software engineering (García-Crespo et al., 2009; Martinho et al., 2010), customer relationship management (García-Crespo et al., 2010a, b), consultancy (Colomo-Palacios et al., 2010), learning environments (Fernández-Breis et al., 2009), biomedical data access (García-Sánchez et al., 2008), human resources management (Soto-Acosta et al., 2010) to cite the most recent efforts.

In order to reach the concept described by the semantic web, it is necessary for resources to be associated with metadata. One mechanism for associating such metadata is annotation. In particular, we may wish to annotate resources with semantic metadata that provides some indication of the content of a resource (Bechhofer et al., 2002). Semantic web annotations go beyond familiar textual annotations about the content of the documents; they formally identify concepts and relations between concepts in documents and the annotations are intended primarily for use by machines (Uren et al., 2006). Unlike an annotation in the normal sense, a semantic annotation must be explicit, formal, and unambiguous: Explicit makes a semantic annotation publicly accessible, formal makes a semantic annotation publicly agreeable and unambiguous makes a semantic annotation publicly identifiable (Ding et al., 2006).

Semantic web annotation contributes two types of additional benefits when compared to plain metadata annotation: Enhanced information retrieval and improved interoperability (Uren et al., 2006). In spite of the advantages of semantic annotation, according to Benjamins et al. (2008), a potential barrier to the uptake of semantic web technology is the effort required to mark up web information with semantic annotations. The same authors indicate that the efforts to obtain semantic annotation seem to be slowly breaking the chicken-and-egg problem that tainted the overall semantic web effort.

In this scenario, the current focus of semantic web research is recasting the Web by providing methods to add semantics to data, manually or automatically, thereby moving the Web toward easier machine processing (Benjamins et al., 2008). Annotation tools may fall into several types: manual, semi-automatic or automatic. Early semantic annotation systems, for example, Annotea (Kahan and Koivunen, 2001), SHOE (Heflin and Hendler, 2001), COHSE (Bechhofer and Goble, 2001), Melita (Ciravegna et al., 2002) and OntoMat-Annotizer (Handschuh et al., 2001), mainly rely on manually submitted semantic descriptions. The necessity to create faster and more accurate recommendation mechanisms has motivated more recent progress on semi-automatic annotation mechanisms, such as MnM (Vargas-Vera et al., 2002), SCORE (Sheth et al., 2002), OWLIR (Shah et al., 2002) and CREAM (Handschuh and Staab, 2003) or including initiatives in automated mechanisms such as SemTag and Seeker (Dill et al., 2003), Armadillo

(Chapman et al., 2004), PANKOW (Cimiano et al., 2005) or KIM (Kiryakov et al., 2004), later taken over by SEKT project, and more recently, efforts like APOLDA (Wartena et al., 2007). Liu and Li (2009) provide a brief summary of annotation technology based on ongologies and propose a Chinese semantic annotation method. Recent efforts related to semantic annotation are shown in Kim et al (2010) and Scheiber et al. (2008).

Currently, the problems generated by the inefficiency and slow speed of manual annotation exist alongside the imperfections characteristic of automatic annotation systems, as shown in recent research on automatic annotations based on semantic web services (Argüello-Casteleiro et al., 2007; Tamma, 2010) or the Wordnet ontology (Yang and Lin, 2010).

However, the reality is that in annotation environments with reduced vocabularies, whose objective is not to annotate entire pages but concrete elements, the subjectivity of the annotator is limited. Thus performance increases, but at the cost of not using automatic and semiautomatic approaches.

Undoubtedly, in the efforts to ensure that the semantic web can exploit more accurate information retrieval mechanisms, the annotation of information is not the only interesting field of research for information processing. Another trend is to use multiple ontologies to satisfy a user search query (Mena et al., 2000). The work of Bhogal et al. (2007) contains one of the most important initiatives in the incorporation of ontologies of specific domains for the improvement of queries. The importance of the use of multiple ontologies is given, because in many practical cases, it is impossible to describe the meanings of web resources without using multiple ontologies (Wang et al., 2004). From a merely technical viewpoint, and uniting annotation and information retrieval in the same field of study, relevant research works have been published whose aim is to tackle the problem of the requirement for annotation of contents using more than one ontology. These efforts (Wang et al., 2004) present the creation of a Bridge as a technical solution, defined as a specific ontology, which can be created and maintained conveniently, and is effective in the multi-ontologies based semantic annotation. Other works propose the integration of ontologies as a technical solution for multi-ontology annotation (Dong and Li, 2006). Some recent efforts also include the annotation with multiple ontolgies in observational datasets as Bowers et al. (2010) or complex biomedical applications as Gennari et al. (2010).

SEAN

Here, the SEAN framework is presented by describing its conceptual model, followed by a discussion of the different annotation methods, as well asoutlining the architecture and, finally, giving relevant information

regarding its implementation.

Highly accurate closed domains (HACD)

The SEAN conceptual model is based on the notion of highly accurate closed domains (HACD). HACD are well-described conceptually specific domains where a few set of concepts are expressive enough to encompass the universe of discourse. A concept is then one of the building blocks of an ontology, particularly those gathering most of the semantics of the ontology. These domains are conceptualizations with a very limited set of fundamental concepts with its relations.

Once the HACD is defined, the set of specific concepts must fulfill a number of ontological constraints, such as having a particular set of relationships, axioms and well-established ontological foundations. However, this is not critical in this case, since one of the most remarkable challenges in recent Information Engineering research is precisely attaining significant benefit from reduced shared information domains by adding semantics as machine-readable annotation. This exploits the benefits of, for example, faceted search versus pure syntactic keyword search. This is the case for the domains which are described subsequently which follow. With this add-on, the annotation of content-specific information becomes possible and the content becomes not only machine-readable, but also machine-processable.

Annotation method

The objective of annotation using SEAN is to describe both the elements which have been entitled projects, as well as the products (documents, code…) associated with each of these, for example, in the case of a software development project. The search for an element can be realized both by persons related to the projects, as well as by any of the projects which implement SEAN (Colomo-Palacios et al., 2008; Gómez-Berbís et al., 2008; García-Crespo et al., 2011). This is a necessity for the sharing of information implies that the annotations should be directed towards the activities of people. Based on this premise, cognitive annotation (Caussanel et al., 2002; Azoau et al., 2004; Lortal et al., 2005) is the most suitable option for annotating both projects and/or products. Any element produced during the project, for example, a requirement, a diagram, source code or documentation, is considered a product.

The objective of the annotation is to provide advanced searches and facilitate the retrieval of information. In an environment oriented to multiple projects, SEAN performs annotation at two levels:

1. On the first level, the annotation refers to the general characteristics of the project. This annotation aids the localization of projects based on these characteristics. Examples of these characteristics could be the type of project (research, development…), the size, or the topics related to the subject of the project. By extending this ontology, new characteristics and products could be included.
2. On the second level, within each project is a set of products in generated, titled products. These products should also be annotated to allow the recuperation of information within each of the projects.

With regard to the annotation of products, techniques for automatic annotation based on natural language can be useful in text documents. Dealing with an application in which the elements to annotate are generic, and given that these elements which are text based can be generated by a heterogeneous group of people, a method is proposed for manual annotation based on ontologies. Both projects as well as products, which constitute the platform for annotation, can be from a wide spectrum of functional domains. Additionally, the users of the system can be from distinct countries and cultures. Due to this circumstance, it is essential that the annotation is based on a common vocabulary. SEAN implements this common vocabulary as two groups of ontologies. On the one hand, an application ontology which describes the different products that can be associated with a project, as well as their, while on the other hand, a domain ontology which will relate the products with the terms of the domain to which the project belongs. The domain ontology provides the common concepts which can be used to describe each of the elements generated. The steps to follow for the annotation of the products are given in the following (Figure 1):

1. Creation of a project. Having created a project, its principal characteristics are defined and are annotated based on the application ontology.
2. Definition of products and related products. The products of the projects and the relations between them are defined using the application ontology. After having created a new product, its characteristics and its related products are defined.
3. Definition of the key words of the domain. Each product has an associated set of key words which relate it with the domain to which the project belongs. The user will select the concepts for each product, based on the domain ontology.

The annotation of the products related to the project is carried out in two phases. Firstly, the moment at which the product is created within the management system, the general characteristics are assigned according to the definition provided by the application ontology. Secondly, once the product is created, additional data is selected based on the domain ontology to which the product
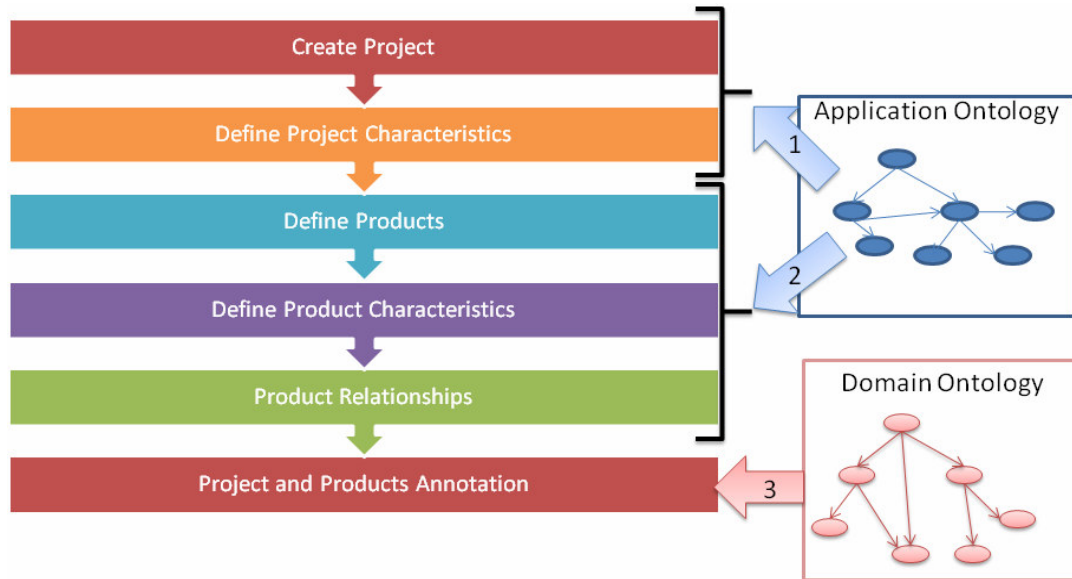
Figure 1. SEAN annotation process.

corresponds.

As previously indicated in the 'State of the art', the slow speed of manual annotation coexists with semiautomatic annotation by using Natural Language Processing (NLP) techniques. Concretely, manual annotation can constitute a bottleneck within the process of the management of information. To avoid this problem, SEAN, that uses manual annotation, proposes two types of actions. Firstly, the limitation of the elements to annotate by means of the use of application ontologies, which clearly determine the number and class of products to annotate, as well as the relations among them. And secondly, with the aim of speeding up the process, the annotation of products as a unique and indivisible element is proposed, instead of annotating their content. For example, to annotate a product (a document) about the description of the architecture of a software application, instead of annotating each of the relevant paragraphs of the document, the semantic annotation will refer to the document as a unique entity, alluding to the unique characteristics which relate the document and its content to the project to which it belongs (by means of the application ontology) and to its domain, by means of the domain ontology. The combination of the application ontology with the domain ontology restricts the number of elements to annotate, as well as the number of annotations to perform, with which sufficient speed is achieved to use manual annotation. With the proposed annotation method, four main advantages are achieved:

1. The number of products and concepts are restricted by well-defined ontologies allowing manual annotation.
2. A project is characterized by a set of attributes based on the application ontology, and a set of keywords based on the domain ontology.
3. Each product related to a project is defined by the application ontology. As previously mentioned, a product could be any element produced during the project, for example, a requirement, a diagram, source code or documentation. The product could be related to other products of the project.
4. Each product has a set of keywords related with the domain of the project.

Architecture

Here, the SEAN architecture, a three layer software architecture, which partitions the functionality of the system into Graphical User Interface (GUI), Business Logic and Persistence and Storage Systems level is presented. Each level has a different functionality to deal with the various challenges SEAN faces when annotating HACDs by means of semantic technologies. The final architectural approach is a tailor-made value-added technological solution, which addresses the aforementioned challenges and provides a basis for the implementation which will be shown subsequently. The SEAN architecture is composed of a number of components, depicted in Figure 2.

The different components, without specifically focusing on the software layer where they belong is detailed in 'USE CASE'. This is not necessary since the three functionalities are well defined and have a commonly shared and used pattern:

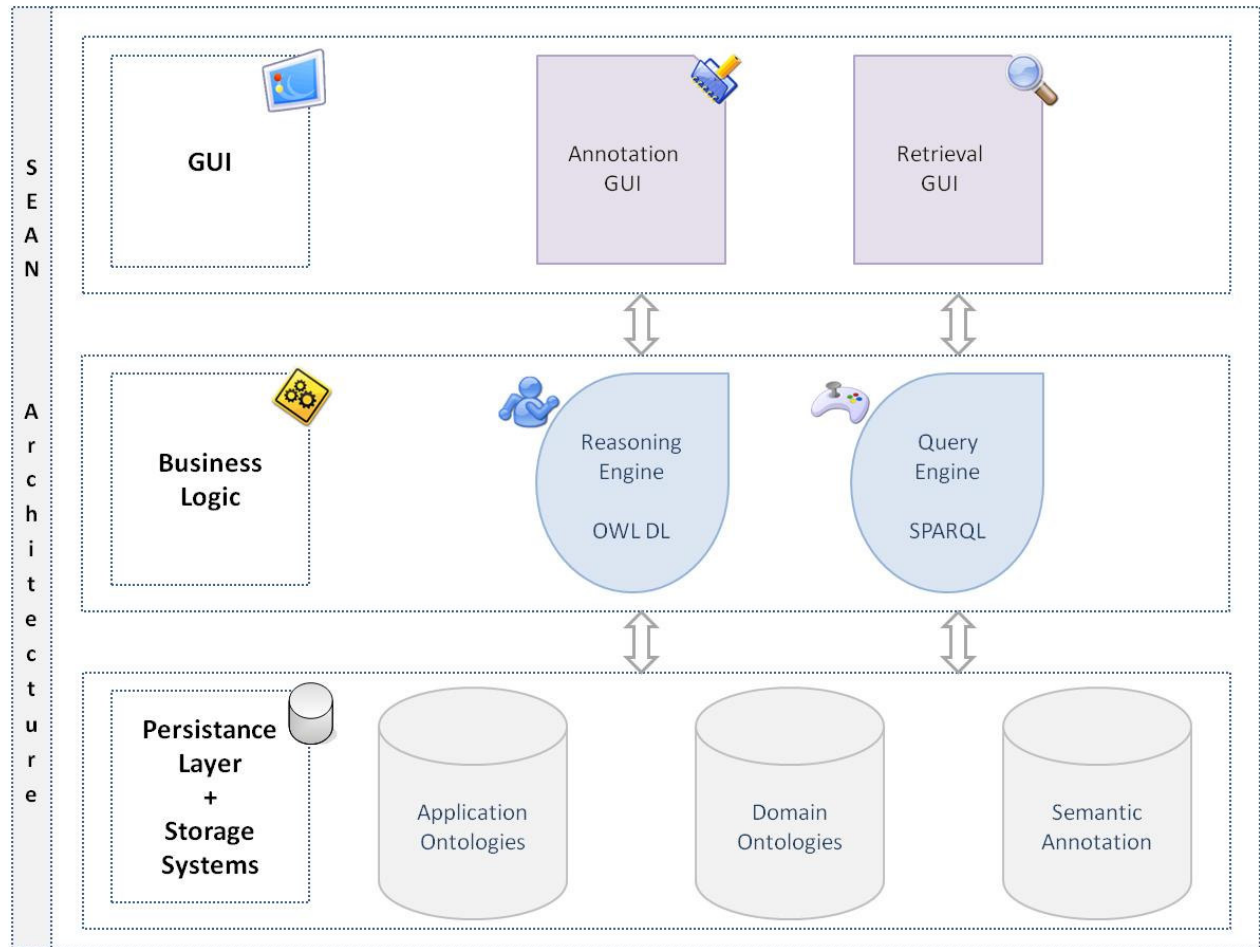1. Annotation GUI: This component interacts with the user, providing a set of graphical elements to annotate

Figure 2. SEAN architecture.

the resources by means of semantic annotations based on Application Ontologies and Domain Ontologies. Annotation GUI has been enhanced with the possibilities offered by the AJAX technology in Java environments. The ontologies are represented in a tree view in order to clearly identify the concepts and their relationships and improve the usability of the system.

2. Retrieval GUI: This constituent offers semantic annotation retrieval functionality for the user, grounded on both the reasoning engine and the query engine. In the former, retrieval is envisaged as location of a subset of concepts by means of description logics subsumption. In the latter, the retrieval is provided by SPARQL definitions to find, manage and query RDF triples following a particular criterion. This GUI is also based on the AJAX technology.

3. Reasoning engine: This component derives facts from a knowledge base, reasoning about the information with the ultimate purpose of formulating new conclusions. In the SEAN framework, it consists of an OWL description logics based reasoner, such as the Renamed ABox and Concept Expression Reasoner (RACER). It uses

subsumption to find sets and subsets of annotations based on logical constraints.

4. Query engine: The query engine component uses the SPARQL RDF query language to make queries into the storage systems of the back end layer. The semantics of the query are defined not by a precise rendering of a formal syntax, but by an interpretation of the most suitable results of the query. SEAN stores mostly RDF triples or OWL DL ontologies, which also present RDF syntax. The JENA framework has been employed in addition to the RACER reasoned in order to implement and optimize the retrieval of information.

5. Semantic annotation, application ontologies and domain ontologies repositories: These three components are semantic data store systems that enable ontology persistence, querying performed by the business logic layer components and offer a higher abstraction layer to enable fast storage and retrieval of large amounts of OWL DL ontologies, together with their RDF syntax. This maintains a small footprint and a lightweight architecture approach. An example of such a system could be the OpenRDF Sesame RDF Storage system or the Yet

Another RDF Storage System (YARS), which deal with data and legacy integration. Jena framework has been used due to it been employed in the other layers o the architecture.

The SEAN architecture is a self-contained, loosely coupled open architecture which allows the use of a wide range of software technologies for its implementation. However, the "use case" focuses on describing a particular implementation, that is, the ESACAKE (Gómez et al., 2008) application which follows the SEAN architectural paradigm and includes current state of the art semantic technologies.

USE CASE

Here, the use of SEAN with a real world case study scenario which demonstrates the contributions of the system is illustrated. Two distinct projects will be considered. Firstly, a software development project being realized in Galway, Ireland. The objective of the project is to construct a mobile financial advisor. This project employs the platform Social Global Repository (SGR) (Colomo-Palacios et al., 2008). This tool enables project members in Galway to adopt a framework that supports European Space Agency (ESA) methodology, and that additionally that allows the querying and sharing of software artifacts (requirements, modules…) among SGR users, both internal and external to the company. In the second place, another company, located in New Orleans, is developing a financial system, for which it uses ProLink (Gómez-Berbís et al., 2008) as support for the management of information in the project.

Thus, the members of Project 1 use the application ontology corresponding to the ESA methodology and a financial domain ontology (García-Manotas et al.'s (2010) ontology) for the annotation of the concepts of the domain. On the other hand, the members of Project 2 use DOAP [1] (Description Of A Project) for the annotation of the characteristics of the project, and the previously mentioned financial ontology to annotate the concepts dealt within this ontology which are relevant to the financial system.

Obviously, both the DOAP ontology and the financial ontology are used for querying, since this allows finding a number of hidden added-value relationships among the project, for example, addressing how many people are required for the project, which domain it belongs to, location and a set of Dublin core-like information.

Both applications, Prolink and SGR have used the services provided by SEAN for the annotation and storage of information, which resides in the persistence layer. The ontologies used have been previously uploaded in the persistence layer, with the objective that both projects can be annotated using the application and domain ontologies.

Once the information has been stored, the retrieval capacities of SEAN can be put into practice. In particular, using the query engine component of the Business Logic layer, any concept belonging to either of the ontologies can be consulted from any application which incorporates SEAN. It results as evident that, using two distinct application ontologies, the results may be disparate from the information point of view, but undoubtedly the annotation of the domain ontology may demonstrate valid and interesting results. For example, consider that the development Project boss in Galway wants to know details of financial application projects. Using the query engine, he introduces a financial concept, for example "fixed term deposits", and the retrieval GUI returns information relating to the project in New Orleans. Taking into account the type of information available in Prolink, It is clear that the result from the point of view of information concerns DOAP data and is not related to ESA. Without doubt, the combined use of the domain and application ontologies enables distinct users to benefit from multi-ontology based queries and obtain useful data (the URL for downloading the software).

EVALUATION

Research design

With the objective of obtaining feedback concerning the work realized, an evaluation was carried out by means of the application of a questionnaire. The questionnaire was applied after the subjects had utilized the ESACAKE (Gómez et al., 2008) tool, under two different architectures. In the first of these, the users applied ESACAKE in its original form and design without using the new SEAN platform and using a manual annotation tool such as those developed at the SWAN project, where semantic annotation is done from a manual, not even semi-automatic perspective, and secondly, they used a modified version of the ESACAKE tool to apply the SEAN platform. The questionnaire had a double objective. In the first place, it was determined if the new SEAN platform improved the annotation, retrieval and sharing capacities of the data present when compared with ESACAKE, while in the second place, the user was requested to provide information about the extension capacities of SEAN when compared with other known platforms.

The questionnaire was composed of three sections. In the first place, the subject was required to provide identification data: Age and gender. Secondly, the users were asked about the different perceptions they had about the use of SEAN in relation with ESACAKE. Thirdly, the subject was asked about the capacities of SEAN in relation to the extension of its use in the context of other applications. With the objective that the task would be standard for the entire set of subjects, a case was designed which the subjects could carry out without difficulty, given their knowledge of ESACAKE. In relation to the first of the objectives, the task consisted of the insertion of two software requirements, their semantic annotation, and a semantic search using the platform. Regarding the second objective, in relation to extensibility, the users were instructed to perform a search related to the requirement previously introduced, but input to another tool in which SEAN was implemented, SGR (Colomo-Palacios et al., 2008). Once the search was carried out, the subjects were asked to compare the capacities of integration between platforms, in this case between ESACAKE and SGR, through the sharing of

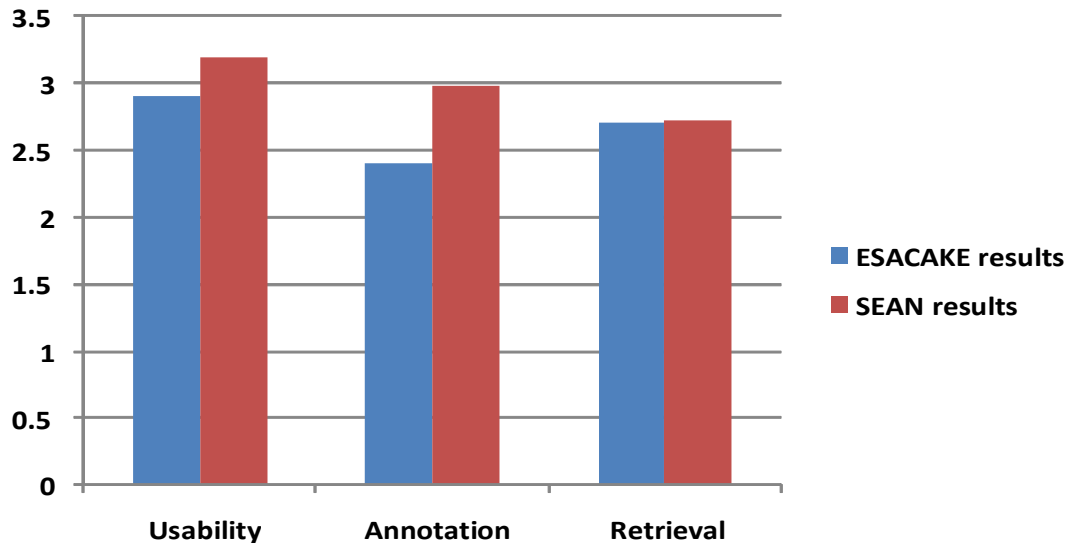---

[1] http://trac.usefulinc.com/doap

Figure 3. Comparison between ESACAKE and SEAN.

ontologies was realized by SGR.

Thus, after the completion of the task, the questionnaire was administered to the subjects, who completed it individually. Subsequent to filling out the identification elements of the questionnaire, it was required that the user responded to two blocks of three questions relative to usability, semantic annotation and retrieval capacities of ESA-CAKE without SEAN, and, on the other hand, of ESACAKE with SEAN incorporated. The responses to these questions were coded using a Likert scale ranging from 1 to 4 points, with the following values. 1: Limited, 2: Regular, 3: Good, and 4: Very Good.

Lastly, the user was asked about the advantages of the combination of SEAN with more than one platform, and concretely, the aim was for the user to provide his opinion about the cross-searches which SEAN allows and the sharing of information between different platforms. The responses to the questions were closed in the same format as before, and coded using the same scale previously described.

In this sense, the choice of an even scale is guaranteed, as well as providing the required correspondence with the scale adopted in the curricular initiative, to avoid what has been termed "central tendency error". This error, which is defined as the reluctance on the part of respondents to give extreme responses (Yu et al., 2003) is limited by obliging the subject to select from a range of values which do not contain a central value.

`The objectives from the point of view of the statistical method was to establish if significant differences exist between using and not using SEAN, and on the other hand, elucidate whether according to the subjects, the extensibility which SEAN provides is considered important.

Sample

The sample was composed of students in the final year of the Computer Science degree of the University Carlos III. These students use the ESACAKE tool to carry out the drawing up of user requirements in the course "Software Engineering III". The sample was composed of 17 women (32%) and 35 men (68%), with an average age of 25.6. Although this population might not completely reflect future users, most studies in the literature have used academics to provide queries and judge relevance (Morrison, 2008).

RESULTS

The results of the surveys, which were realized using printed copies, were subsequently coded. On the one hand, Figure 3 depicts the results relative to the questionnaire relative to the comparison between ESACAKE and SEAN. On the other hand Figure 4 shows the results relative to the evaluation of the extensions of SEAN. In Tables 1 and 2, the average and standard deviation of the responses offered by the students are shown in relation to the questionnaire applied, and the two groups of questions formulated, respectively. Fundamentally, a cross search (Table 2) is a more complex search which is performed over several search subsets. Information sharing (Table 2) is a parameter that measures to which extent information is accessible for the others.

One of the objectives of the study is to determine whether differences in use exist depending on using SEAN or not. In order to perform such an analysis, the statistical method Student's t-test (comparison of two means) was used to carry out one-way between-groups analysis of variance. The level of statistical significance was set at 0.05. The results of the test indicate that the annotation element presents significant differences between both populations, indicated by the statistical value (t (52) = -4.88, p < 0.05). This circumstance implies that, from a statistical point of view, there is a difference between the annotations of both architectures, and considering the average, the annotation is improved in SEAN. Additionally, the usability of the integrated SEAN
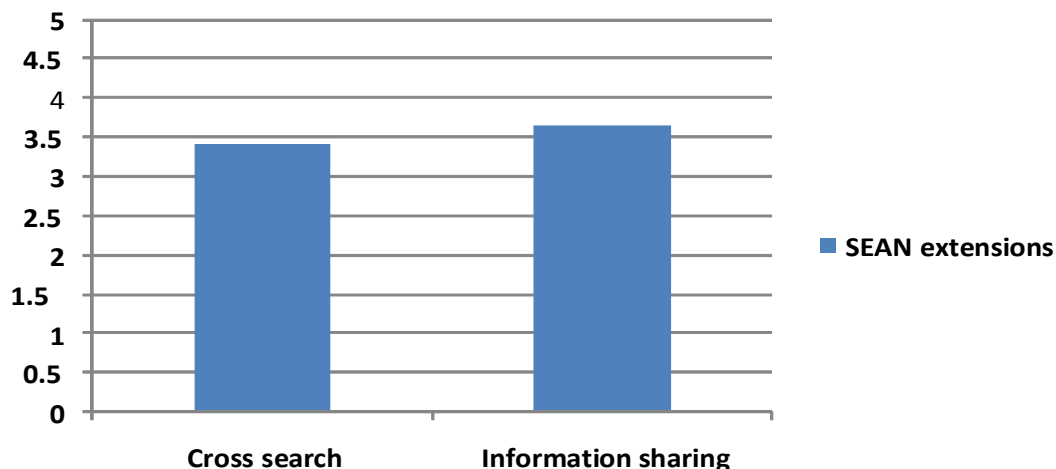
Figure 4. Evaluation of SEAN extensions.

Table 1. Non SEAN Vs. SEAN statistical results.

	ESACAKE		SEAN	
	Average	Std. deviation	Average	Std. deviation
Usability	2.91	0.49	3.19	0.55
Annotation	2.40	0.57	2.98	0.66
Retrieval	2.70	0.51	2.72	0.60

Table 2. SEAN extension statistical results.

	Average	Std. deviation
Cross search	3.42	0.57
Information sharing	3.66	0.52

platform presents significant differences between both populations (t (52) = -2.77, p < 0.05). This circumstance may be a result of the fact that the integration of SEAN in the platform produces improvement in the use of the application for the user.

DISCUSSION

Considering the results demonstrated, two conclusions can be drawn. In the first place, the platform presents improved usability, annotation and retrieval characteristics, evidenced by the versions of the tools which do not yet incorporate SEAN. In the second place, with regard to the information provided by the users, the integration capacities which SEAN provides are very high.

In relation to the first group of conclusions, the opinions of the users have a limited agreement. This pattern may

be due to the fact that the GUI of the annotation tools in ESACAKE has been only recently designed, and even though in the SEAN version they are improved, the level of improvement is still only moderate. In all of the aspects analyzed, the use of SEAN improves previous features on average, but undoubtedly presents some major standard deviations, from which it can be determined that due to the dispersion of opinions, there is less cohesion of agreement. In relation to the statistical differences between the usability, annotation and retrieval criteria, the Students' t-test reveals that both annotation and usability preset significantly improved characteristics in SEAN.

With respect to the second group of conclusions, the value of SEAN as an agglutinator of semantic information and a platform which allows shared access to information between platforms has been confirmed. In particular and especially consulting the scores, which present an average of 3.66 points and an adjusted standard

deviation of 0.52, the users consider the capacities for the sharing of information very interesting.

In summary, this results show how the implementation of SEAN has statistically improved the previous ESACAKE implementation in the tested areas. We have also found that new features provided by SEAN have been accepted by the users of the case of study. These results also show that more improvements could be made in future research and work, for example in the GUI.

Conclusions

This paper introduces SEAN, a novel framework for semantic annotation of HACD. The framework is not focused on finding perfect solutions for very complex wide open domains, since the complexity grows dramatically, and previous state of the art approaches show the tremendous difficulty of annotating such domains. Instead of this, the current approach is committed to contributing an efficient solution by means of manual semantic annotation for well-defined, consensus shared, accurate closed domains. The ESACAKE application was implemented and tested as a proof-of-concept implementation for software development project requirements based on the ESA standard. The current approach was studied, compared and evaluated in relation to similar systems. Three aspects demonstrate the advantages of SEAN: The potential for well-defined domains semantic annotation, consensus sharing and minimal semantic complexity applied to a given domain. Although, the annotation method still requires improvement and it could be complemented by NLP based semi-automatic annotation, the researchers believe that the work on the implementation, test and evaluation of SEAN justifies further development of annotation of concrete domains. Fundamentally, the main contribution of this work is also that SEAN can be used in a different set of platforms. SEAN was designed with a clear user focus and the only possibility to learn and feedback from annotation is extending to the widest user base, what implies trying to abstract from the software limitations of a particular platform, by means of a Web-based user interface and also with a significant set of domain ontologies for a multi-domain strategy.

Regarding future work, this paper has focused on finding the best, possibly approximate, solution for manual semantic annotation in particular domains. However, the strengths of the present approach could be harnessed using NLP based techniques for automatic or semi-automatic annotation that, probably without taking the human out of the loop, would efficiently improve the scalability and performance of the system. Another issue to be addressed in the future is the application of the SEAN architecture in other domains, for example for enabling the knowledge sharing in other areas like development of new product of services in small and medium enterprises (Ebrahim et al., 2010).

REFERENCES

Alani, H, Hall W, O'Hara K, Shadbolt N, Szomszor M, Chandler P (2008). Building a Pragmatic Semantic Web. IEEE Intell. Syst., 23(3): 61-68.

Argüello-Casteleiro M, Abusa M, Fernandez-Prieto MJ, Brookes V, Abanda FH (2007). A web services-based annotation application for semantic annotation of highly specialised documents about the field of marketing. In: Meersman, R., Tari, Z. (eds.) Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part I. Berlin, Heidelberg: Springer-Verlag,, pp. 1135-1152.

Azouaou F, Desmoulins C, Chen W (2004). Semantic Annotation Tools for Learning Material. In: Arroyo, L., Dicheva, D. (Eds.) Proc. of Workshop on Applications of Semantic Web Technologies for Adaptive Educational Hypermedia Systems, pp. 359-364.

Battré D (2008). Caching of intermediate results in DHT-based RDF stores. International Int. J. Metadata Semant. Ontol., 4(3): 183-195.

Bechhofer S, Carr L, Goble C, Kampa S, Miles-Board T (2002). The Semantics of Semantic Annotation. In Proceedings of CoopIS/DOA/ODBASE, pp. 1152-1167.

Bechhofer S, Carr L, Goble C, Kampa S, Miles-Board T (2002). The Semantics of Semantic Annotation. In: Meersman, R., Tari, Z. (Eds.), Proceedings of CoopIS/DOA/ODBASE, Berlin / Heidelberg, Springer, pp. 1152-1167.

Benjamins VR, Davies J, Baeza-Yates R, Mika P, Zaragoza H., Greaves, Gómez-Pérez, JM, Contreras, J Domingue J, Fensel D (2008). Near-Term Prospects for Semantic Technologies. IEEE Intell. Syst., 23(1): 76-88.

Berners-Lee T, Hendler J, Lassila O (2001). The semantic web. Sci. Am., 284(5): 34-43.

Bhogal J, Macfarlane A, Smith P (2007). A review of ontology based query expansion. Inform. Process. Manag., 43(4): 866-886.

Bowers S, Cao H, Schildhauer M, Jones M, Leinfelder B, O'Brien MA (2010). Semantic annotation framework for retrieving and analyzing observational datasets. In Proceedings of the third workshop on Exploiting semantic annotations in information retrieval. New York, NY, USA: ACM, pp. 31-32.

Cardoso J (2007). The Semantic Web Vision: Where are We?. IEEE Intell. Syst., 22(5): 22-26.

Caussanel J, Cahier JP, Zacklan M, Charlet J (2002). Cognitive Interactions in the Semantic Web. In: Frank, M., et al. (eds) Proceedings of the WWW2002 International Workshop on the Semantic Web, Hawaii, May 7, 2002.

Chapman S, Dingli A, Ciravegna F (2004). Armadillo: harvesting information for the semantic web. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY: ACM, pp. 598-598.

Cimiano P, Ladwig G, Staab S (2005). Gimme' the context: context-driven automatic semantic annotation with C-PANKOW. In Proceedings of the 14th international conference on World Wide Web. New York, NY: ACM, pp. 332-341.

Ciravegna F, Dingli A, Petrelli D, Wilks Y (2002) User-system cooperation in document annotation based on information extraction. In Gómez-Pérez, A., Benjamins, V.R. (Eds.), 13th International Conference on Knowledge Engineering and Management (EKAW 2002), LNCS 2473. Berlin / Heidelberg: Springer, pp. 122-137.

Colomo-Palacios R. García-Crespo Á, Soto-Acosta P, Ruano-Mayoral M, Jiménez-López D (2010). A case analysis of semantic technologies for R&D intermediation information management. Int. J. Inf. Manage., 30(5): 465-469.

Colomo-Palacios R, Gómez-Berbís JM, García-Crespo A, Puebla-Sánche I (2008). Social Global Repository: using semantics and social web in software projects. Int. J. Knowl. Learning, 4(5): 452-464.

Davies J, Lytras MD, Sheth AP (2007). Semantic-Web-Based Knowledge Management. IEEE Int. Comp., 11(5): 14-16.

Dill S, Eiron N, Gibson D, Gruhl D, Guha R, Jhingran A, Kanungo T,

McCurley KS, Rajagopalan S, Tomkin A, Tomlin JA, Zien JY (2003). A Case for Automated Large Scale Semantic Annotations. Web Semant., 1(1): 115–132

Ding Y (2010). Semantic Web: Who is who in the field — a bibliometric analysis. J. Inf. Sci., 36(3): 335-356.

Ding Y, Embley DW, Liddle SW (2006). Automatic Creation and Simplified Querying of Semantic Web Content: An Approach Based on Information-Extraction Ontologies. In: Mizoguchi, R., Shi, Z., Giunchiglia, F. (eds) Proceedings of the First Asian Semantic Web Conference (ASWC'06). Berlin Heidelberg: Springer, pp. 400-414.

Dong A, Li H (2006). Multi-ontology Based Multimedia Annotation for Do-main-specific Information Retrieval. In: Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous and Trusworthy Computing (SUTC' 06). Los Alamitos, California: IEEE Computer Soc. Press, pp. 158-165.

Durguin JK, Sherif JS (2008). The semantic web: a catalyst for future e-business. Kybernetes, 37(1): 49-65.

Ebrahim NA, Ahmed S, Taha Z (2010). SMEs; Virtual research and development (R&D) teams and new product development: A literature review. Int. J. Phys. Sci., 5(7): 916-930

Euzenat J (2002). Research challenges and perspectives of the Semantic Web. IEEE Intell. Syst., 17(5): 86-88.

Fensel D, Munsen MA (2001). The Semantic Web: A Brain for Humankind. IEEE Intell. Syst., 16(2): 24-25.

Fensel D (2002). Ontologies: A silver bullet for knowledge management and electronic commerce. Berlin: Springer.

Fensel D, van Harmelen F, Horrocks I, McGuinness DL, Patel-Schneider PF (2001). OIL: An ontology infrastructure for the semantic web. IEEE Intell. Syst., 16(2): 38-45.

Fernández-Breis JT, Castellanos-Nieves D, Valencia-García R (2009). Measuring Individual Learning Performance in Group Work from a Knowledge Integration perspective. Info. Sci., 179(4): 339-354.

Ferreira N (2010). Social Networks and Young People: A Case Study. Int. J. Hum. Capital Inf. Technol. Prof., 1(4): 31-54.

García R (2010). Using the Rhizomer Platform for Semantic Decision Support Systems Development. Int. J. Decision Support Syst. Technol., 2(1): 60-80.

García-Crespo A, Colomo-Palacios R, Gómez-Berbís JM, García-Sánchez F (2010a). SOLAR: Social Link Advanced Recommendation System. Future Gener. Comp. Sys., 26(3): 374-380.

García-Crespo A, Colomo-Palacios R, Gómez-Berbís JM, Mencke M (2009). BMR: Benchmarking Metrics Recommender for Personnel issues in Software Development Projects. Int. J. Comput. Intell. Syst., 2(3): 257-267.

García-Crespo A, Colomo-Palacios R, Gómez-Berbís JM, Ruiz-Mezcua B (2010b). SEMO: a framework for customer social networks analysis based on semantics. J. Inf. Technol., 25(2): 178-188.

García-Crespo Á, Gómez-Berbís JM, Colomo-Palacios R, García-Sánchez F (2011). Digital Libraries and Web 3.0. The CallimachusDL Approach. Comput. Hum. Behav., doi:10.1016/j.chb.2010.07.046

García-Manotas I, Lupiani G, García-Sánchez F, Valencia-García R (2010). Populating Knowledge Based Decision Support Systems. Int. J. Decision Support Syst. Technol., 2(1): 1-20.

García-Sánchez F, Fernández-Breis JT, Valencia-García R, Gómez JM, Martínez-Béjar R (2008). Combining Semantic Web Technologies with Multi-Agent Systems for Integrated Access to Biological Resources. J. Biomed. Inform.., 41(5): 848-859.

Gennari JH, Neal ML, Galdzicki M, Cook DL (2010) Multiple ontologies in action: Composite annotations for biosimulation models. J. Biomed. Inform., DOI: 10.1016/j.jbi.2010.06.007.

Gómez BJM, Colomo PR, García CA, Ruiz MB (2008). ProLink: a semantics-based social network for software projects. Int. J. Technol. Manage., 7(4): 392-405.

Gómez JM, Mencke M, Chamizo J, Colomo R, García-Crespo A (2008). EsaCake: A Semantic Software Environment for Sharing Software Projects Knowledge Based on the ESA Software Methodology. In: A. Mellouk, J. Bi, G. Ortiz, D. Chiu, M. Popescu (eds.) Proceedings of the Third International Conference on Internet and Web Applications and Services (ICIW 2008). Los Alamitos, California: IEEE Computer Society Press, pp. 535-540.

Gruber TR (2008). Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Web Semant., 6(1): 4-13.

Gruber TR (1993). A translation approach to portable ontology specifications. Know. Acquis., 5(2): 199-220.

Guha R, McCool R, Miller E (2003). Semantic Search. In: WWW '03 Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary. New York, NY: ACM, pp. 700-709.

Handschuh S, Staab S (2003). Cream: Creating metadata for the semantic web. Coput. Netw., 42(5): 579-598.

Handschuh S, Staab S, Maedche A (2001). CREAM – creating relational metadata with a component-based, ontology-driven annotation framework. In: First International Conference on Knowledge Capture (K-CAP 2001). New York, NY: ACM, pp. 76-86.

Heflin J, Hendler JA (2001). Portrait of the Semantic Web in Action. IEEE Intell. Syst., 16(2): 54-59.

Hendler J (2008). Web 3.0: Chicken Farms on the Semantic Web. Comp., 41(1): 106-108.

Horrocks I (2008). Ontologies and the Semantic Web. Commun. ACM, 51(12): 58-67.

Kahan J, Koivunen MR (2001). Annotea: an open RDF infrastructure for shared web annotations. In: WWW '01 Proceedings of the 10th international conference on World Wide Web. New York, NY: ACM, 623-632.

Kim HL, Decker S, Breslin JG (2010) Representing and sharing folksonomies with semantics. J. Inf. Sci., 36(1): 52-72.

Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D (2004). Semantic Annotation, Indexing, and Retrieval. Web Semant., 2(1): 49-79.

Lassila O, Hendler J (2007). Embracing "Web 3.0". IEEE Internet Comp., 11(3): 90-93.

Liu Y, Li ZZ (2009) Research of Semantic Annotation Technology Based on Domain Ontology. In: Proceedings of the 2009 Second International Workshop on Computer Science and Engineering - Volume 02. Washington, DC: IEEE Comp. Soc., 358-361.

Lortal G, Lewkowicz M, Todirascu-Courtier A (2005). AnT&CoW, a tool supporting collective interpretation of documents through annotation and indexation. In: Dieng R., Matta N. (eds.) Proceedings of Workshop on Knowledge Management and Organizational Memories - IJCAI'05, Workshop, pp. 43-54.

Lukasiewicz,T, Straccia U (2008). Managing uncertainty and vagueness in description logics for the Semantic Web. Web Semant., 6(4): 291-308.

Markoff J (2006). Entrepreneurs See a Web Guided by Common Sense. The N.Y. Times, 12 Nov. 2006.

Martinho R, Varajao J, Domingos D (2010). Using the semantic web to define a language for modelling controlled flexibility in software processes. IET Software, 4(6): 396-406.

Mena E, Kashyap V, Illarramendi A, Sheth A (2000). Imprecise Answers in Distributed Environments: Estimation of Information Loss for Multi-Ontology based Query Processing. Int. J. Coop. Inf. Syst., 9(4): 403-425.

Mika P (2005). Flink: Semantic Web technology for the extraction and analysis of social networks. Web Semant., 3(2/3): 211-223.

Mikroyannidis A, Theodoulidis B (2010). Ontology management and evolution for business intelligence. Int. J. Inf. Manage., 30(6): 559-566.

Morrison PJ (2008). Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. Inform. Process. Manage., 44(4): 1562-1579.

Naeve A (2005). The Human Semantic Web Shifting from Knowledge Push to Knowledge Pull. Int. J. Semantic Web Inf. Syst., 1(3): 1-30.

Scheir P, Lindstaedt SN, Ghidini C (2008). A Network Model Approach to Retrieval in the Semantic Web. Int. J. Semantic Web Inf. Syst., 4(4): 56-84.

Schreiber G, Amin A, Aroyo L, Assem M, de Boer V, Hardman L, Hildebrand M, Omelayenko B, van Osenbruggen J, Tordai A, Wielemaker J, Wielinga B (2008). Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Web Semant., 6(4): 243-249.

Shadbolt N, Hall W, Berners-Lee T (2006). The semantic web revisited. IEEE Intell. Syst., 21(3): 96-101.

Shah U, Finin T, Joshi A (2002). Information retrieval on the semantic web. In: Proceedings of the 11th international conference on Information and knowledge management. New York, NY: ACM, pp.

461-468.

Sheth A, Bertram C, Avant D, Hammond B, Kochut K, Warke Y (2002). Managing semantic content for the web. IEEE Internet Comp., 6(4): 80-87.

Smith B (2003). Ontology. An Introduction, In Floridi, L. (ed.), Blackwell Guide to the Philosophy of Computing and Information. Oxford: Blackwell, pp. 155-166.

Soto-Acosta P, Casado-Lumbreras C, Cabezas-Isla F (2010). Shaping human capital in software development teams: the case of mentoring enabled by semantics. IET Software, 4(6): 445-452.

Studer R, Benjamins VR, Fensel D (1998). Knowledge engineering: Principles and methods. Data Knowl. Eng., 25(1-2): 161-197.

Tamma V (2010) Semantic Web Support for Intelligent Search and Retrieval of Business Knowledge. IEEE Intell. Syst., 25(1): 84-88.

Uren VS, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Cira-vegna F (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semant., 4(1): 14-28.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F. (2002). MnM: ontology driven semi-automatic and automatic support for semantic markup. In: Gómez-Pérez, A., Benjamins, V.R. (Eds.) 13th International Conference on Knowledge Engineering and Management (EKAW 2002), LNCS 2473. Berlin / Heidelberg: Springer, pp. 213-221.

Vossen G, Lytras MD, Koudas N (2007). Editorial: Revisiting the (Machine) Semantic Web: The Missing Layers for the Human Semantic Web. IEEE Trans. Knowl. Data Eng., 19(2): 145-148.

Wang P, Xu BW, Lu JJ, Li YH, Jiang JH (2004). Bridge ontology: A multi-ontologies-based approach for semantic annotation. Wuhan Uni. J. Nat. Sci., 9(5): 617-622.

Warren P (2006). Knowledge Management and the Semantic Web: From Scenario to Technology. IEEE Intell. Syst., 21(1): 53-59.

Wartena C, Brussee R, Gazendam L, Huijsen WO (2007). Apolda: A Practical Tool for Semantic Annotation. In: Proceedings of the 18th Int. Conference on Data-base and Expert Systems Applications (DEXA '07). Washington, DC: IEEE Computer Society Press, pp. 288-292.

Yang CY, Lin HY (2010). An automated semantic annotation based-on Wordnet ontology. In: Ko, F., Na, Y. (eds.) Sixth International Conference on Networked Computing and Advanced Information Management (NCM), pp. 682-687.

Yu JH, Albaum G, Swenson M (2003). Is a central tendency error inherent in the use of semantic differential scales in different cultures? Int. J. Mark. Res., 45(2): 213-228.