Semantic Information Retrieval: a return on experience

R. Carolina Medina-Ramírez*

Abstract—**In previous works, we have presented the advantages of using a domain ontology and annotations on information retrieval as well as the translation problems between languages with different expression semantic levels. In this paper, we extend our previous work, presenting a return on experience and focusing on the viewpoint of the end-user. In fact, we explore the impact and helpfulness of a domain ontology, semantic annotations relying on this ontology and semantic resource descriptions so as to enrich end-user responses extracted from an information retrieval system. A system embodying this approach is presented.**

Keywords: Semantic information retrieval, ontology

1 Introduction

The semantic Web is an extension of the current web in which information is given a well-defined meaning, so as to be accessible and comprehensible not only to humans but also to computers thus enabling computers and cooperation among people [13]. This approach relies on ontologies (information exchange and search), semantic annotations (document content representation) and formal knowledge representation languages (for representing these ontologies and annotations). The ongoing works on this direction have produced several methods, knowledge representations formalisms and tools to annotate and manipulate web resources in a semantic manner. In the last few years, an increasing generation of ontology-guided Information Retrieval systems focused on ontology knowledge representation languages have been proposed(SHOE [7], On2Broker [6], OntoSeek [11], WebKB [12], Corese [4, 5]). They propose an ontology-guided retrieval of annotated documents. Nevertheless, the huge amount of proposed formalisms shows not only the increasing interest of such approaches but also the problems faced when sharing annotations and ontologies. We argue that translation methods are needed to share and re-use knowledge by using languages with different levels of semantic expression. In addition, among the heterogeneous resources belonging for example to a scientific community or to an enterprise, documents (in digital or paper supports) constitute a significant source of knowledge needing to be represented, handled, queried and diffused. Besides, the Web Community is invested in developing new semantic search techniques, but the question of personalizing the interaction with web content is at hand. Web users aim at retrieving resources or services satisfying specific criteria or constraints. They want retrieved resources to be displayed in a personalized format. Particularly, results from desktop search engines are still limited. Typical formats of retrieved documents consist of a list of results containing a set of lines describing the document found. The corresponding description is based on the keywords submitted in the query. Important information such as: document type (journals, proceedings or informal notes), publication date, author names, journal and conferences name are missing in a real response from the web. The presence of such information in the returned results is of relevant importance to select the pertinent document from a specific user query. A much richer expressiveness than simple keywords for describing resources is definitely needed. The goal of this paper is to describe not only the mechanisms for representing document contents for automating certain processing in applied fields such as information retrieval or knowledge management, but also an environment for managing, capitalizing and distributing knowledge into an information retrieval framework. We claim that an effort has to be made in the displaying of results for a better comprehension and transfer of knowledge. The rest of this paper is structured as follows. In Section 2, we briefly depict the framework of this paper by describing the ESCRIRE project. Section 3 presents the semantic elements, annotations and a domain ontology, defined in the ESCRIRE project. In Section 4, we discuss the EscorServer architecture. Section 5 describes our approach for enriching end-user responses. In Section 6 we present some concluding remarks as well as some directions of our work.

2 The ESCRIRE project

The framework of this work was the ESCRIRE project [8], the first goal of which was to compare three knowledge representation formalisms (KR): conceptual graphs (CG), descriptions logics (DL), and object-oriented representation languages (OOR) for querying about docu-

*Universidad Autónoma Metropolitana-Iztapalapa, Electrical Engineering Department, *Redes y Telecomunicaciones* research team, San Rafael Atlixco 186, Col. Vicentina, 09340 Iztapalapa, Mexico Tel/Fax: +52.55.5804.4629/4628 E-mail: cmed@xanum.uam.mx

(Advance online publication: 17 November 2007)

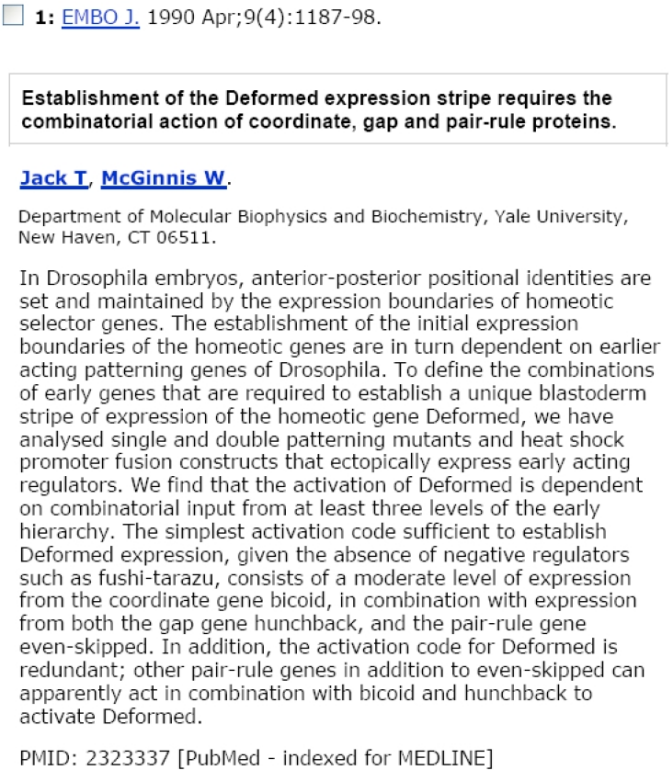
ment contents by relying on ontology-based annotations on document content. This comparison relies on an XML-based pivot expressive language to define the ontology and to represent annotations and queries; it consists of evaluating the capabilities of the three KR formalisms for expressing the features of the pivot language. Each feature of the pivot language is translated into each KR formalism, which is then used to draw inferences and to reply queries. As a first return on experience of this process, we encountered problems during the information (ontology and annotations) exchange (to share and re-use knowledge).We have discussed and underlined in [2] the main problems encountered during the translation among languages with different expressivity semantic levels. The second goal of the ESCRIRE project was the representation and handling of document contents for document retrieval. The corpus chosen for experimenting with ESCRIRE is composed of scientific summaries of articles (abstracts) related to the genetic interactions leading to the segmentation process of the drosophila fly. These abstracts are obtained from the Pubmed database [10]. A test base composed of a set of 4500 abstracts of articles on biology from PubMed with semantic annotations on their contents was used. The format of the response proposed by ESCRIRE was simple; it consists of a list of pertinent documents and the submitted query. In Section 5 we detail our proposed approach for enriching such format.

3 Semantic elements in ESCRIRE project

3.1 Annotations

A formal representation of document content allows to make structured requests and thus to seek and retrieve documents in an efficient manner. With the aim of making accessible and comprehensible such knowledge by a machine, we proposed to describe the document content in a semantic manner. These semantic descriptions are called annotations. In order to describe semantically a document we need to consider two points; the first point consists in choosing the relevant elements so as to represent knowledge formally. This process may be done in a manual or in a semiautomatic fashion. The second point is to find mechanisms to exploit this knowledge in order to spread and to capitalize that knowledge. The abstracts of documents are textual and contain sentences such as: "*... even-skipped can apparently act in combination with bicoid and hunchback to activate Deformed...*".

Some alternatives to represent the content of documents are presented in [9]. These alternatives go from an exhaustive representation of the document to more targeted representations depending on the application that uses these annotations. The approach adopted by the ESCRIRE project was to carry out a targeted representation. We were interested in the genetic interactions between genes, genes and the classes of genes concerned. With this intention, ESCRIRE adopted a "top-down" approach for the analysis of the documents of the corpus. For example, for the sentence "*... even-skipped can apparently act in combination with bicoid and hunchback to activate Deformed ...*" which represents one of the interactions cited in the article shown in Figure 1, we can make the following analysis to detect and obtain its formal representation.



1: EMBO J. 1990 Apr;9(4):1187-98.

Establishment of the Deformed expression stripe requires the combinatorial action of coordinate, gap and pair-rule proteins.

Jack T, McGinnis W.

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511.

In Drosophila embryos, anterior-posterior positional identities are set and maintained by the expression boundaries of homeotic selector genes. The establishment of the initial expression boundaries of the homeotic genes are in turn dependent on earlier acting patterning genes of Drosophila. To define the combinations of early genes that are required to establish a unique blastoderm stripe of expression of the homeotic gene Deformed, we have analysed single and double patterning mutants and heat shock promoter fusion constructs that ectopically express early acting regulators. We find that the activation of Deformed is dependent on combinatorial input from at least three levels of the early hierarchy. The simplest activation code sufficient to establish Deformed expression, given the absence of negative regulators such as fushi-tarazu, consists of a moderate level of expression from the coordinate gene bicoid, in combination with expression from both the gap gene hunchback, and the pair-rule gene even-skipped. In addition, the activation code for Deformed is redundant; other pair-rule genes in addition to even-skipped can apparently act in combination with bicoid and hunchback to activate Deformed.

PMID: 2323337 [PubMed - indexed for MEDLINE]

Figure 1: Document 90214629 resulting from our corpus of work

Level 1: General description: the document refers to the presence of genes that belong to the drosophila;

Level 2: The implied genes belong to classes *primarypair-rule*, *anterior-gap* and *anterior-system*;

Level 3: The representer of class *primary-pair-rule* is the gene named *even-skipped* symbolized by *eve*. In a similar way, the gene named *hunchback* symbolized by *hb* is an instance of the class *primary-pair-rule*. Finally, the gene named *bicoid* symbolized by *bcd* is a representer of the class *anterior-system*;

Level 4: The identified genes have an influence (positive) during the process of segmentation of the drosophila (information related to the field);

Level 5: The *even-skipped* gene activates the *deformed* gene, the *hunchback* gene activates the *deformed*

gene and the bicoid gene activates the deformed gene. Thus, the following code shows the formal representation (in the ESCRIRE language) of those interactions.

```
<esc:relation type="interaction">
 <esc:role name="promoter">
   <esc:objref type="gene" id="eve">
 </esc:role>
 <esc:role name="target">
   <esc:objref type="gene" id="Dfd">
 </esc:role>
 <esc:attribute name="effect">
   <esc:value>activation</ esc:value >
 </ esc:attribute >
</esc:relation>
<esc:relation type="interaction">
 <esc:role name="promoter">
   <esc:objref type="gene" id="hb">
 </esc:role>
 <esc:role name="target">
   <esc:objref type="gene" id="Dfd">
 </esc:role>
 <esc:attribute name="effect">
   <esc:value>activation</ esc:value >
 </ esc:attribute >
</esc:relation>
<esc:relation type="interaction">
 <esc:role name="promoter">
   <esc:objref type="gene" id="bcd">
 </esc:role>
 <esc:role name="target">
   <esc:objref type="gene" id="Dfd">
 </esc:role>
 <esc:attribute name="effect">
   <esc:value>activation</ esc:value >
 </ esc:attribute >
</esc:relation>
```

We considered additional interaction information (if available in the text) concerning the attributes attached to interactions such as the effect, the moment or the localization of the influence.

3.2 A domain ontology

Representing formally the whole content of a document without loosing information is a difficult task [9]. During the development of the ESCRIRE project, we decided to focus on genes, on the genetic interactions during the segmentation process of the fly and the implied gene classes. The entities charged to annotate the document contents are gene classes, genes and interactions among them. Those entities constitute the ESCRIRE ontology and are used to build annotations. In the context of comparing knowledge representation formalisms,

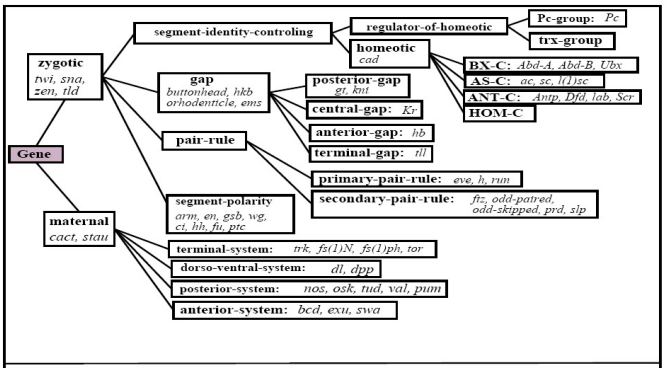


Figure 2: Taxonomy of the classes of genes and genes resulting from the corpus of work.

it was necessary to have objects and relations. The ESCRIRE language is able to describe objects in a document, their attributes, as well as to indicate their membership to classes. Moreover, it makes possible to describe classes and to organize them in a taxonomy. The relations are seen as objects. It is important to remark that the ESCRIRE project separates the ontology (description of the genes and their classes) and the instances (declaration of the interactions between genes). Besides, the ESCRIRE language lets to formulate queries relying on annotations, to represent document content and to define a domain ontology. This language is composed of three sublanguages: (ESC) for ontology and annotations descriptions, (QESC) for queries and (RESC) for result formatting. Nevertheless, the genes and their organization into classes, which are explicitly named in several papers, represent well each time the same objects. The different entities describe a field consensus and the documents refer only to those entities. So, we decided to represent the gene classes, the genes and their taxonomic organization into an ontology since these entities are used as reference in other documents. The taxonomic organization of genes found in the corpus of work is shown in Figure 2. This figure emphasizes in italics the instances of genes.

4 EsCorServer architecture

The information retrieval needs in the Web are presented in different scales in scientific communities, also called corporate Semantic Webs. The framework of the semantic Web can be applied to these communities in order to benefit from that approach. In particular, among the heterogeneous resources belonging, for example, to a scientific community or to a company, documents (in electronic or paper supports) constitute a significant source of knowledge that needs to be represented, handled, queried and diffused. With the aim to capitalize and diffuse the knowledge on genetic interactions in the documentary memory, we propose EsCorServer. This system is a document server that handles, shares and capitalizes

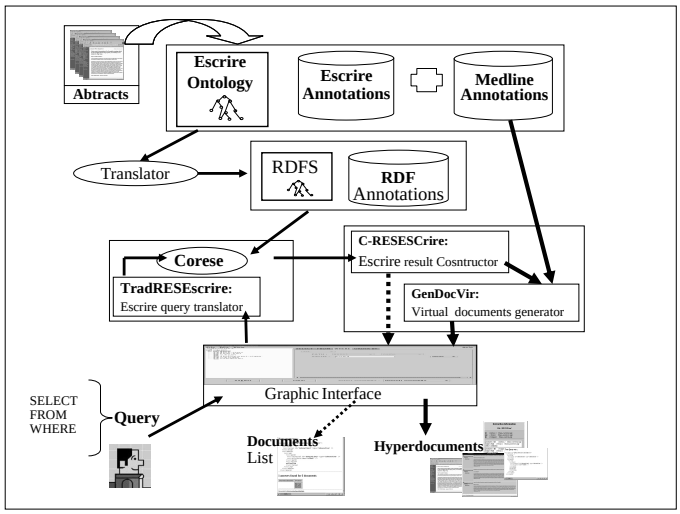


Figure 3: EsCorServer Architecture.

explicit knowledge (document content and data) from a specific domain (Drosophila melanogasters gene interactions) for information retrieval. EsCorServer is based on an ontology-guided information retrieval, semantic annotations of domain articles abstracts, PubMed descriptions and adaptive hypermedia techniques. The heterogeneous aspects of this documentary memory reside on the nature of its resources and on the representation format of its document contents.

Figure 3 shows the EsCorServer architecture. The main element is an interface for introducing, translating and displaying results from a query. The translation and retrieving mechanisms are described in [2, 3]. The GenDocVir module is charged to generate on-demand the interaction information document. We describe this document in Section 5. However, the technology has evolved considerably since the design stage of the EsCorServer. In the area of adaptive hypermedia, the issue of authoring adaptive hypermedia systems is still one of the most important research issues in this area[1].

5 Enriched end-user response approach

We use ontology and resources description to enrich the response given to the user. We can easily access to information annotated by exploiting the Corese semantic search engine[5]. For instance, given the next query in natural language "...To show documents in which the effect of the interactions is the activation ...", the following code represents the last query in ESCRIRE language.

```
<esc:query url=http://escrire.inrialpes.fr
     xmlns:esc=http://escrire.inrialpes.fr/>
<esc:select/>
 <esc:from>
  <esc:relvar id=interaction1 type=interaction/>
 </esc:from>
 <esc:where>
 <esc:eq>
  <esc:path>
<esc:relvarref id=interaction1 type=interaction/>
    <esc:attribute name=effect/>
  </esc:path>
   <esc:value>activation</ esc:value >
 </esc:eq>
</esc:where>
</esc:query>
```

Applying the above query to EsCorServer we obtain the following documents as result: `90015118`, `90214629` and `90292349`. These numbers correspond to the PubMed Identifier.

The enriched end-user response approach shown in Figure 4 consists of creating a hyperdocument composed of the abstracts from documents retrieved by the Corese search engine. This hyperdocument has also links to additional documents: the original document in PubMed, the query made and the interaction informations (created on-demand). The author-s name, publication date, journal and PubMed identifier are included in the hyperdocument as well in order to provide additional useful information .

The document referring to the interaction information is generated on-demand by integrating semantic descriptions of gene interactions (Escrire annotations) and the concepts of a domain ontology. This document contains particular information such as: genes description (scientific gene names, belonging family, activation or inhibition effects, participant genes names of interactions mentioned in the article). Figure 5 shows the interaction information from file `90214629`. It describes four interactions between genes, one of them (in white) has an inhibition effect caused by the gene *fushi tarazu* (ftz) over the gene *deformed* (Dfd). The other three ones (in orange) involve the promoter genes: *eve*, *hb*, *bcd* which produce an activation effect over the gene *deformed* (Dfd).

More specific learning scenarios and profiles must improve the adequacy between the annotation contents and the end-user requests. The innovative aspect of the approach described in this paper and the contribution to the field of adaptive hypermedia documents is the merging of different resource descriptions. This provides robustness to end-user responses to a query as well as to the way of accessing information annotated got from the use of the Corese semantic search engine.

In our experiment, we use a proprietary knowledge representation language (ESCRIRE language) to represent domain ontologies as well as annotations. We found some translation problems while using RDF(S) [3]. In the context of the semantic web retrieval, using languages such

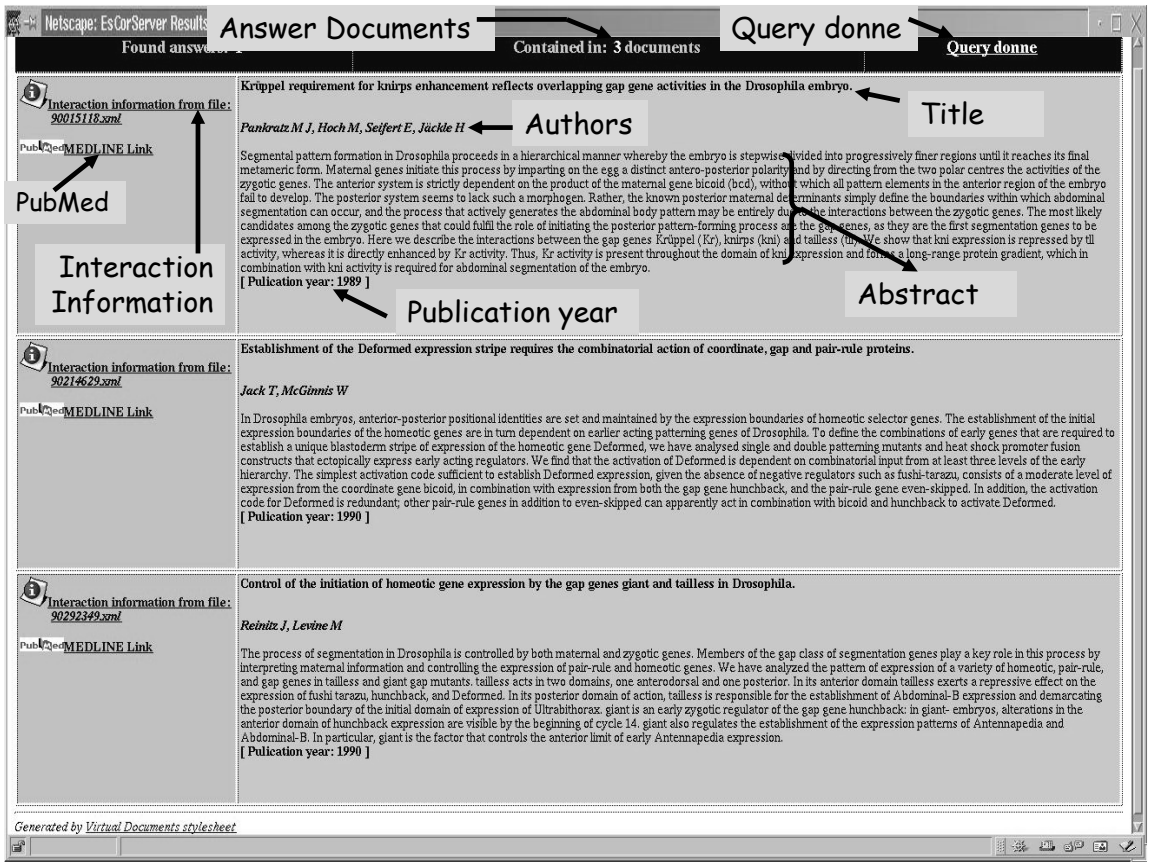


Figure 4: Enriched end-user approach.

Interaction Information
File: 90214629

ftz *(fushi tarazu)* Class= ***secondary-pair-rule***
Dfd *(deformed)* Class= ***ANT-C***
eve *(even-skipped)* Class= ***primary-pair-rule***
hb *(hunchback)* Class= ***anterior-gap***
bcd *(bicoid)* Class= ***anterior-system***

promoter =**ftz** *(fushi tarazu)* Class= ***secondary-pair-rule***
target = **Dfd** *(deformed)* Class= ***ANT-C***
effect = inhibition

promoter =**eve** *(even-skipped* Class= ***primary-pair-rule***
target = **Dfd** *(deformed)* Class= ***ANT-C***
effect = activation

promoter =**hb** *(hunchback)* Class= ***anterior-gap***
target = **Dfd** *(deformed)* Class= ***ANT-C***
effect = activation

promoter =**bcd** *(bicoid)* Class= ***anterior-system***
target = **Dfd** *(deformed)* Class= ***ANT-C***
effect = activation

Figure 5: Interaction information generated on demand corresponding to File 90214629

as RDFS or OWL is recommended so as to model and share the knowledge of a specific user community. From the experience got in this work, we believe that using proprietary languages is not recommended since they are often not compatible with the architecture of the Semantic Web.

We evaluated our prototype within a representative group of experts as well as a group of non-experts in the domain of the Drosophila fly. The results obtained show that using the implicit information in the ontology and in annotations is well suited for the needs of information retrieval. The document that corresponds to the interaction information created on-demand, allows to get a better understanding of the subject for the non-expert people.

6 Conclusion and Future Work

The Web as is used nowadays performs a function in society that transcends its main technical characteristics. It will improve considerably and will help us to manage, integrate and analyze data, as well as to publish and discover documents. However, the single information elements within those documents cannot be handled directly as data. New paradigms are needed to obtain the best benefit from the huge amount of available information on the Web. The semantic Web takes faces these challenges

and enables a continually evolving set of new services. From the experience got from this work, we believe that manual annotation of resources is overwhelming to domain experts or teachers when they are faced to a large amount of resources. So, it is necessary to automate as much as possible the extraction of knowledge from structured format documents. The next challenge is to create the hyperdocument by adding semantic resource descriptions according to user interests and to present them in a manner that facilitates exploration and motivates the user. Preliminary evaluations of our prototype have produced encouraging results. So, our future work will focus on an extension of our prototype to analyze additional results on this direction.

References

[1] De Bra P., Aerts A., Smits D., Stash N. "AHA! Version 2.0 More Adaptation Flexibility for Authors", *Proceedings of the AACE ELearn*, 10/2002

[2] Medina-Ramírez, C., Corby, O., Dieng-Kuntz, R., "A Conceptual Graph and RDF(S) approach for representing and querying document content", *Advances in Artificial Intelligence-IBERAMIA 2002, 8th Ibero-American conference on AI. Ganjo Francisco J., Riquelme J.Cristóbal., Toro M. (Eds.)*. LNCS 2527, Seville, Spain, pp. 121-130, 11/02

[3] Medina-Ramírez, C., Corby, O., Dieng-Kuntz, R., "Querying a heterogeneous corporate semantic Web: A translation approach", *Proceedings of the international workshop on "Knowledge Management through Corporate Semantic Webs". During the EKAW conference*, Singüenza, Spain, pp. 53-63, 10/02

[4] Corby O., Dieng R., Hébert C. "A Conceptual Graph Model for W3C Resource Description Framework", *Proceedings of the 8th International Conference on Conceptual Structures (ICCS)*, LNCS 1867, Darmstadt, Germany, pp. 468-482, 08/00

[5] Corby O., Faron-Zucker C., "Corese: A Corporate Semantic Web Engine", *Proceedings of the WWW2002 Workshop on Real World RDF and Semantic Web Applications*, Honolulu, Hawai, USA, 05/02

[6] Fensel D., Angele J., Decker S., Erdmann M., Schnurr H.P., Staab S., Studer R., Witt A., "On2broker: Semantic-Based Access to Information Sources at the WWW", *Proceedings of the World Conference on the WWW and Internet: WebNet*, pp. 366-371, 10/99

[7] Luke S., Spector L., Rager D., Hendler J.,"Ontology-based Web Agents", *Proceedings of the First International Conference on Autonomous Agents*, pp. 59-68, 10/97.

[8] Al-Hulou R., Corby O., Dieng-Kuntz R., Euzenat J., Medina-Ramírez C., Napoli A., Troncy R., "Three knowledge representation formalisms for content-based manipulation of documents", *Proceedings of the KR 2002 Workshop on Formal Ontology, Knowledge Representation and Intelligent Systems for the World Wide Web (Semweb)*, Toulouse, France, 04/02.

[9] Raphaël Troncy, "Intégration texte-représentation formelle pour la gestion de documents XML", Rapport de Stage de DEA, Université Joseph Fourier, INRIA-Rhône-Alpes, 2000.

[10] Medline database, `http://www.ncbi.nlm.nih.gov/PubMed`, 2002.

[11] Guarino N., Masolo C., Vetere G.,"OntoSeek: Content-Based Access to the Web", *IEEE Intelligent Systems*, V14, N3, pp. 70-80, 10/99

[12] Martin P., Eklund P.W.,"Knowledge Retrieval and the World Wide Web", *IEEE Intelligent Systems.*, V15, N3, pp. 18-25, 05/00

[13] Shadbolt N., Berners-Lee T., Hall W., "The Semantic Web Revisited", *IEEE Intelligent Systems*, V21, N3, pp. 96-101, 05/06