A Semantic Web Portal with HLT Capabilities

Florence Amardeilh(*,**), Thomas Francart(**)
florence.amardeilh@paris4.sorbonne.fr, thomas.francart@mondeca.com

(*)LaLICC - Université de Paris IV - UMR CNRS
96, Bd Raspail Paris, 75006, France

(**)Mondeca
3, cité Nollez Paris, 75018, France

Keywords:
Semantic Web, semantic annotations, information extraction, knowledge base enrichment.

Mots clefs :
Web Sémantique, annotations sémantiques, extraction d'information, enrichissement de bases de connaissance.

Palabras clave :
Web Semántica, anotaciones semánticas, extracción de información, enriquecimiento de bases de conocimiento.

Abstract

In this paper we will present a new form of Semantic Web portal using Human Language Technologies (HLT). Our system provide the means to annotate documents with metadata, populate a knowledge base according to the corresponding domain ontology and most of all, update the linguistic resources with all the new extracted information in order to improve the performance of the entire system. As a consequence, it will assist the Web communities, and among them the competitive intelligence workers, to create domain-centric Semantic Web portals. The final user will be able, through the application's interfaces, to vizualise the knowledge base data, to formulate intelligent and complex queries and at least, to publish the returned results. We would like to point out the fact that the entire platform is Semantic Web-compliant as it is based on its standards (XML, XTM, RDF(S) and OWL).

Résumé

Dans cet article, nous allons présenter une nouvelle forme de portail Web Sémantique utilisant des Traitements du Langage Naturel (TLN). Notre système fournit les moyens d'annoter les documents avec des metadonnées, d'enrichir une base de connaissance dépendamment de l'ontologie du domaine correspondant, et surtout de mettre à jour les ressources linguistiques utilisées avec les nouvelles informations extraites afin d'améliorer la performance du système dans son ensemble. Par conséquent, ce nouveau système permet d'assister les communautés sur le Web, et notamment les personnes travaillant dans le domaine de la veille scientifique et économique, à créer des portails Web Sémantique centrés sur le domaine d'application. L'utilisateur final sera capable à travers les interfaces de l'application de visualiser les données de la base de connaissance, de formuler des requêtes intelligentes et complexes et enfin de publier les résultats trouvés. Il est important de noter que la plate-forme décrite dans la suite de cet article est conforme aux standards et langages du Web Sémantique (XML, XTM, RDF(S) et OWL).

1. Introduction

According to Tim Berners-Lee [BER 98], the vision of a Semantic Web consists of making the actual Web comprehensible and thus exploitable by the machines. To achieve that goal, the data must be annotated and structured semantically by adding sense and knowledge through the annotation of the resources by semantic tags [KAT 02]. Those tags act as many clues for the machines to interpret, process and combine the information almost like humans do [LU 02]. As a consequence, human users could exploit these semantically tagged resources to query, share, access, or publish them and thus work more effectively [LAU 02].

Because of the problems coming from the annotation of the existing documentary corpora, from the productivity and quality needs of the created annotations, it is essential for the Semantic Web success to have methods allowing the semi-automatic production of annotations from unstructured documents, i.e. to extract the knowledge of a domain of application and to populate a knowledge base with the instances of the concepts, as well as their properties and the relations between these concepts. These concepts and relations are defined by the domain ontology and the knowledge could then be managed and exploited by non-expert final users, such as documentalists or competitive intelligence workers.

To construct these needed methods for the Semantic Web, we naturally think of using Human Languages Technologies (HLT). Actually, the linguistic technologies can have a major impact on knowledge management and especially on Semantic Web communities to construct operational solutions for users. We believe that a strong collaboration between these two areas of research will greatly improve the comprehension of the Web by the machines and will become the basis for a future generation of intelligent tools for the Web.

First, this paper presents the reasons of our interest in the HLT to develop a Semantic Web portal (section 2). Then we describe the actual architecture of our system and its main components (section 3). In the following part, we will present our research work based on the mapping between the linguistic tools and the domain ontology featuring an example in the competitive intelligence domain (section 4). At last, we will go through a short overview of related projects (section 5) before concluding and discussing on future work (section 6).

2. Towards a Semantic Web integrated portal using HLT

On the one hand, research on knowledge representation, developed in the Knowledge Management field, has a strong tradition in domain specific knowledge description. Those techniques allow the process of this knowledge by the machines. On the other hand, Semantic Web is based on knowledge representation systems, especially by the use of ontologies, and the comprehension and exploitation by the machines of the documentary resources, either coming from the Internet or from the companies' information systems. Even if more and more documents are created dynamically from databases, the unstructured textual information is still dominating. And the Human Language Technologies community is precisely specialized in knowledge representation from textual documents.

Thus it seems natural to us to get the best of the methods and tools originated from this community of research so they can, thanks to computational means, linguistically process texts. Among the existing linguistic technologies, we will focus our research on Information Extraction (IE). IE is composed of several linguistic methods to find and extract pertinent information according to a domain from a textual documentary corpus.

The extracted information is mainly composed of named entities such as proper nouns (of persons, organizations, locations, etc.) or numbers (amounts, percentage, measures, etc.) and dates (absolutes and relatives). The notion of "named entities" (NE) has been defined through the different Message Understanding Conferences[1] (MUC). The purpose of these MUC conferences was to measure the precision and efficiency of the developed technologies for extracting predefined named entities and semantic relations (or "scenarios") between these named entities on semi-structured textual documents.

Thanks to the functionalities offered by the Human Language Technologies, adaptive solutions to the Semantic Web needs can be implemented, such as:

- the semi-automatic constitution of vocabularies/terminologies of a domain from a representative documentary corpus;
- the semi-automatic enrichment of a knowledge base by the named entities and their semantic relations extracted from the textual documents;
- the semantic annotation of these documents.

We are working on the architecture of a Semantic Web portal integrating existing linguistic tools. According to the above solutions, we focused our research on how to use Information Extraction methods to annotate textual documents and to populate a knowledge base. To improve productivity and quality of human indexation, the new information extracted and added to the knowledge base will serve to enrich the linguistic resources. As a consequence, the system is able to reuse its components' results in order to improve its overall performance. The portal architecture is detailed in the next section of this document.

3. The ITM Semantic Web Portal

Our Semantic Web portal solution is based on Mondeca's Intelligent Topic Manager™ (ITM) tool. ITM is a knowledge management software and development platform, with automatic knowledge acquisition capabilities. It is based on the Semantic Web standards. The terminological and ontological resources [BOU 03] in ITM are based on the OWL recommendation [OWL 04]. They are developed with Protégé 2.0[2] and imported in ITM. Ontologies and thesaurus can be further edited in ITM if changes are required, or they can even be created from scratch in the semantic knowledge portal. The underlying knowledge base is based on the XTM (XML Topic Map) language, which defines sets of topics and associations. This language allows a greater flexibility in knowledge representation, particularly when modeling complex n-ary semantic relations.

3.1 Architecture of the ITM portal

ITM integrates a built-in semantic knowledge portal, with intuitive user-interfaces. This portal provides four key features, summarized on Fig. 1:

1. **Query**. A user can query the knowledge base taking into account the constraints issued from the domain ontology and the thesaurus. The querying interfaces are adapted to the user's profile and its information need.
2. **Navigation**. A user can browse the content of the knowledge base through a textual and/or a graphical interface. The user can navigate through the representations of entities and semantic relations.
3. **Publication**. A user can organize a set of documents and/or knowledge corresponding to the result of a precedent query into a publication. The user can choose the output format of this publication depending on his/her needs: XML, HTML, PDF, TXT, etc.
4. **Edition**. A user can edit any item in the knowledge base, the ontology or the thesaurus or add new items, through ontology-controlled interfaces.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
[2] http://protege.stanford.edu/index.html

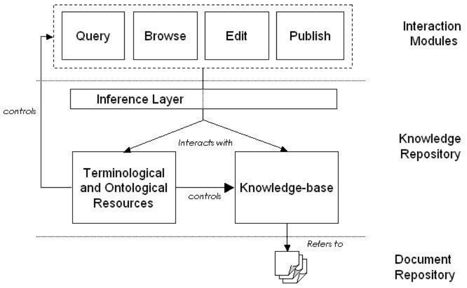


Fig. 1: Overall architecture of the ITM platform

As summarized on the precedent figure, the ontology constrains the population of the knowledge base, and the interaction modules. The interaction between these modules and the knowledge repository goes through an inference layer, providing class heritage, transitive relations processing, etc. The knowledge base points to the actual documents, accessible through a URL or stored in an external content management system.

3.2 Characterizing an ontology in the ITM portal

ITM provides a meta-ontology, defining the core concepts used in the platform, such as `Class, Attribute`, or the `Class-subclass` relationship. This meta-ontology controls the structure of all the specific domain ontologies that will be deployed in the portal. Informally, an ontology in ITM defines a hierarchy of classes along with the possible types of attributes and relations of their instances. Formally, an ontology in ITM is composed of [MAE 03]:

1. A set of classes C. C represents the possible classes of the instances in the domain.
2. A hierarchy H. Classes are related by the irreflexive, acyclic, transitive relation H, ($H \subset C * C$). H(C1,C2) means that C1 is a subclass of C2. Note that it is allowed for a class not to be part of the hierarchy.
3. A set of attribute types D. D is the set of possible types of data that can be attached to the instances in the domain.
4. A set of association types A. A is the set of possible types for the associations between the instances in the domain.
5. A set of role types R. R is the set of possible types for the roles played by instances in the domain.
6. An attribute-constraint function AC: $C \rightarrow P(D)$, defining for each class the list of possible types of attributes on instances of this class.
7. A set of association constraints RC. RC is a relation (p,q,r), with $p \in C$, $q \in R$, $r \in A$. A relation constraint (p,q,r) expresses the fact that an instance of class p can play a role of type q in an association of type r.

3.3 Characterizing a knowledge-base in the ITM portal

Every knowledge base in ITM is controlled by one or more ontologies, specified when creating it. Thus the possible classes of instances and types of relations are the one defined in these ontologies. Informally, a knowledge base in ITM is a set of instances connected to each other in a semantic network of associations and roles. Every instance has a class, and every associations and roles have a type. For example, one could model the statement “Mr X is the CEO of company Y”, by the network: “Mr X [class Person] is playing a role of type employee in an association of type employment, where CEO [class function] is playing the role of position, and where the company Y [class Company] is playing the role of employer”.

More formally, a knowledge base in ITM is composed of [MAE 03]:

1. A set of instances I. Every element of I has a class $c \in C$.

2. A set of attributes K. Every element of K has a type d ∈ D.
3. An attribute function Ik: I → P(K). Ik defines for each instance i ∈ I the list of attributes of i.
4. A set of associations B. Every element of B has a type a ∈ A.
5. A set of roles S. Every element of S has a type r ∈ R.
6. A "role-instance" function Ri: S → I.
7. A "role-association" function Ra: S → B. (Ri and Ra together expresses the fact that a role connects an instance to an association).

The knowledge base can be populated manually by the end-user through web forms. The user selects in the ontology the class of object he wants to instantiate, and a web-form is generated with the authorized attributes and relations on instances of this class. Thus the information required to create the new item is dependent on the class selected by the user. Nevertheless, the manual population of a knowledge base has its drawbacks: it is time-consuming, error-prone and user dependant even if controlled by the ontology. It has consequences on semantic annotation with regards to the user productivity, the information quality and the document processing frequency. For all those reasons, we decided to improve the ITM Semantic Web portal by integrating a set of HLT to help the user to annotate the documents and to populate the knowledge base.

4 The HLT Contribution to the Semantic Web Portal

As shown in the following figure, the HLT contribution can be effective on both document annotation and knowledge base enrichment. On the one hand, document annotation is the addition of semantic tags (metadatas, descriptors from a thesaurus, named entities, etc.) to the textual document. This document can be stored elsewhere on an external content management system and shared with other applications. On the other hand, the knowledge base enrichment is the population of a knowledge base by the information contained in the document (new descriptors to add to the thesaurus, new named entities, semantic relations, etc.). The semantic tags and the knowledge base are constrained by the domain ontology and the terminological resources.

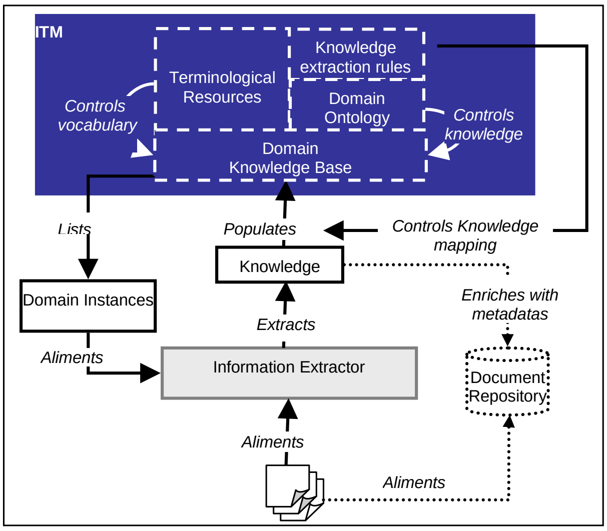


Fig. 2. Our Semantic Web Portal architecture with the HLT component.

Consequently, a fine-grained mapping has to be defined between the domain ontology, the knowledge base and the linguistic tools. This system will be able to parse the entire document and not only its metadata tags. It will provide suggestions of annotations to the user thanks to the linguistic tools. The user will have to validate the extracted information before saving the annotations and populating the knowledge base. It is a semi-automatic process.

In the next sections of this document, we will refer to a competitive intelligence ontology defined in one of our client applications. We mapped the extraction patterns with some of its concepts: named entities such as `Company` and `Product`, and semantic relations between those named entities like `Product_Adoption`, `Company_Acquisition`, `Partners`, etc.

4.1 The linguistic tools

Linguistic analysis is performed by the Insight Discoverer™ Extractor [GRI 01] developed by the company Temis. The information extractor implements the finite-state transducer method after the documents have been parsed through a tokenizer, a sentence splitter, a lemmatizer and a part-of-speech tagger. The finite-state transducer method defines a set of extraction patterns, each of them describing the ways a concept can be described in a textual document. The concept usually represents a named entity or a semantic relation and can be reused to define other concepts in cascade. These extraction patterns are combined with linguistic resources such as dictionaries and thesaurus in order to improve the extraction results. These linguistic resources are represented as lists of terms that belong to the applied domain, like all the main company names and their aliases in the Telecom area (see Fig. 3). This sort of grammar has proven to be applicable for various linguistic tasks and have traditionally been used for IE and Named Entity Recognition (Cf. MUC conferences).

```
<?xml version = '1.0'?>
<component>
<concept name="~CompanyPart">
    <concept name="~CompanyPart.name">
        ;;Orange
        ;;Equant
        France / Telecom
        Swiss / General
        (Hence)? / Nabisco
        (Pastificio)? / Gazzola
        3s-informatique
        @tlantic
        A\. / Moksel
;;        ABB () in sigle
        AT / & / T
        ;;AT&T
        Aachener / und / Münchener
        Abase
        Abbott
        Abside
        Accor
... </concept></concept>
</component>
```

Fig. 3 Companies dictionnary example

4.2 Mapping the extraction patterns to the domain ontology

The process of mapping should conform to the following requirements:

1. **Ease of use**. The mapping process involves experts from different fields (domain expert, linguistic expert and knowledge modelization expert) and is not a straightforward task. Thus the chosen solution should be easily understood by the three parties and should permit an iterative mapping process.
2. **Independence between the ontology structure and the linguistic extraction structure**. Using natural language processing to populate a knowledge-base must not add new constraints on the way the ontology is designed, or on the format of the linguistic extraction.
3. **Capacity to evolve**. The system must be able to take into account the evolutions of both the ontology and the linguistic tool.

4. **Completeness**. The system must be able to retrieve every information given by the linguistic extractions.
5. **Standardization**. The system must not be dependant on the linguistic tool being used.

Here is an example of an extraction result on the competitive intelligence corpora:

```
/CI Extraction(10692,50, Suresnes, France - November 20, 2002 -
Dassault Systemes (Euronext Paris # 13065, DSY.PA, Nasdaq: DASTY) to-
day announced the acquisition of Knowledge Technologies International
(KTI) in an all-cash transaction)
   who(10692,14, Dassault Systemes)
      /actor(10692,14, Dassault Systemes)
   /buying acquisition(10719,3,the acquisition of)
   whom(10723,6,KTI)
      /actor(10723,6,KTI)
when(10730,12,November 20, 2002)
```

Fig. 4. Extraction result example: a company acquisition

The solution we designed to map the linguistic extraction to the ontology concepts is based on XPath[3]. XPath is a language of navigation in information trees, especially in XML. Though not strictly speaking XML, we can clearly see in Fig. 4 that this extraction result has a tree-based format, thus supporting XPath expressions.

In order to achieve the mapping of the ontology and the linguistic extraction, we add a new layer of information to the ontology, namely the knowledge extraction rules (see Fig. 2). These rules are Xpath expressions attached to elements of the ontology we want to instantiate: when a node in the linguistic extraction match a knowledge extraction rule, the corresponding ontology concept is instantiated. We give a sample of the competitive intelligence ontology used along with the knowledge extraction rules associated with each concept of this ontology. Note that the ontology by itself is independent of any linguistic extraction.

Ontology concept	**Associated knowledge extraction rule**
“Company acquisition” is an association type ∈ A.	CI Extraction [buying acquisition] *(rule 1)*
“acquiring” is a role type ∈ R.	CI Extraction/who *(rule 2)*
“acquired” is a role type ∈ R.	CI Extraction/whom (*rule 3*)
“Company” is a class ∈ C	CI Extraction/who/actor or CI Extration/whom/ actor (*rule 4*)
(“Company”; “acquired”; “Company acquisition”) and (“Company”; “acquiring”; “Company acquisition”) are two triples ∈ RC[4].	*No rule associated*

When parsing the information tree using a breadth-first method, the following events will be triggered in order :

1. The node “CI Extraction” is encountered, with a child “buying acquisition”, thus corresponding to rule 1: an association of type “Company acquisition” is instantiated.
2. The node “who” is encountered, with a parent “CI Extraction”, thus corresponding to rule 2: a role of type “acquiring” is instantiated.
3. The node “whom” is encountered, with a parent “CI Extraction”, thus corresponding to rule 3: a role of type “acquired” is instantiated.
4. The node “actor” under the node “who” is encountered, thus corresponding to the rule 4: an instance of “Company” is created. The name of this instance is taken from the text value of this node, `Dassault Systemes`.

[3] http://www.w3.org/TR/xpath

[4] These 2 triples mean that an association of type “Company acquisition” can have two roles, “acquired” and “acquiring”, both played by instances of “Company”.

5. Similarly, the node “actor” under the node “whom” is encountered, rule 4 applies, and another instance of “Company” named KTI is created.

Finally, the roles “acquired” and “acquiring” are tied to the association “Company acquisition”, and respectively to Dassault Systemes and KTI. This is determined from the context of execution.

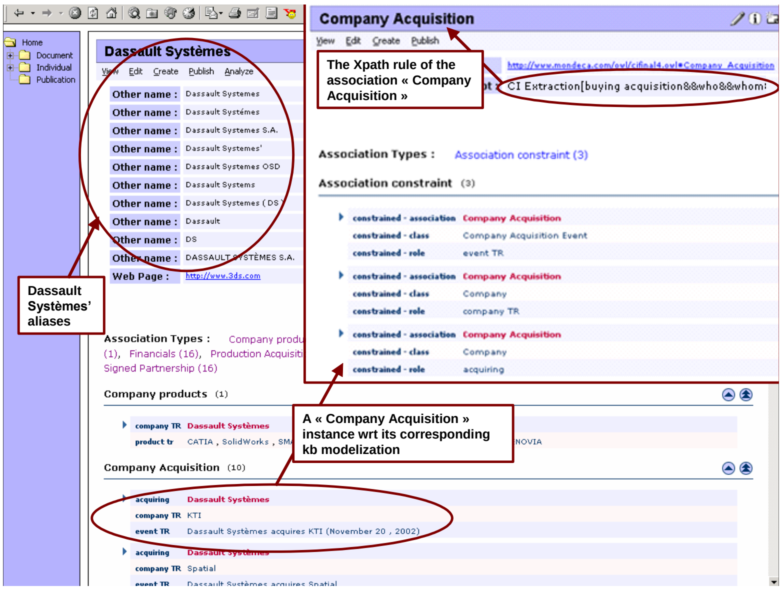
This solution clearly meets the requirements stated above, as it is separated from the ontology, thus allowing great flexibility and independence. It is also standard and not tied to a specific tool, easy to use, and complete, since any part of a tree can be reached by a XPath expression. Through this method, not only linguistic extractions, but also other types of tree-based documents could be integrated into an ontology-controlled knowledge base, especially XML structured or semi-structured documents.

4.3 The Semantic Annotation Process

Document annotation.
Each time a new document needs to be published in the system, it is firstly annotated by the linguistic tools. The information extractor locates the named entities and the semantic relations based on its extraction patterns. Then, according to the terminological resources and the concepts extracted, the system is able to deduce a set of descriptors from the thesaurus and of named entities for the document. These are suggestions provided to the user that has to validate. Then the document is annotated and stored in an external content management system.

Knowledge base enrichment.
Once the extraction has been made, the extracted information is compared to the domain ontology thanks to the mapping explained above. At the same time that the user proceeds to the validation of the document annotations, he also validates the other information extracted that will populate the knowledge base. It is mainly about new descriptors that will become candidates for the thesaurus, new named entities (generally “guessed” by the information extraction patterns) and the semantic relations. In the validation screen, the user can check the suggestions proposed by the system and create new instances in the knowledge base, modify recorded instances (to add aliases or attributes for example) or validate information. In
Fig. 5, the previous semantic relation Company_Acquisition between Dassault Systemes and KTI has been validated and recorded in the ITM semantic knowledge base. In the upper-right window is shown the XPath rule for the semantic relation Company_Acquisition.

Commentaire [T1]: peut-etre pas necessaire de repeter ?



Dassault Systèmes

Other name :	Dassault Systemes
Other name :	Dassault Systémes
Other name :	Dassault Systemes S.A.
Other name :	Dassault Systemes'
Other name :	Dassault Systemes OSD
Other name :	Dassault Systems
Other name :	Dassault Systemes (DS)
Other name :	Dassault
Other name :	DS
Other name :	DASSAULT SYSTÈMES S.A.
Web Page :	http://www.3ds.com

Dassault Systèmes' aliases

Association Types : Company produ (1), Financials (16), Production Acquisiti Signed Partnership (16)

Company products (1)

company TR	Dassault Systèmes
product tr	CATIA , SolidWorks , SM NOVIA

Company Acquisition (10)

acquiring	Dassault Systèmes
company TR	KTI
event TR	Dassault Systèmes acquires KTI (November 20 , 2002)
acquiring	Dassault Systèmes
company TR	Spatial
event TR	Dassault Systèmes acquires Spatial

A « Company Acquisition » instance wrt its corresponding kb modelization

Company Acquisition

The Xpath rule of the association « Company Acquisition »

http://www.mondeca.com/owl/cifinal4.owl#Company_Acquisition

CI Extraction[buying acquisition&&who&&whom

Association Types : Association constraint (3)

Association constraint (3)

constrained - association	Company Acquisition
constrained - class	Company Acquisition Event
constrained - role	event TR
constrained - association	Company Acquisition
constrained - class	Company
constrained - role	company TR
constrained - association	Company Acquisition
constrained - class	Company
constrained - role	acquiring

Fig. 5. Competitive Intelligence example through the semantic knowledge portal

Linguistic resources update.

As the users should not constantly be requested to validate the results obtained by the extractions, the system will automatically update the linguistic resources with the new information extracted and validated by the user. Indeed, the dictionaries of the linguistic tool will be completed with the list of the "guessed" entities that have been validated as new ones, the manually added entities and the already existing entities whose name or aliases have been modified/completed. To complete that task, the system must take every classes defined in the ontology as representing a type of entity (such as `Person`, `Companies`, `Organisation`, `Product`, etc.) where a new instance can be found. The system will extract from the knowledge base those instances for each class of entity constituting an extraction concept. The system will list the concepts and their instances by their names and aliases according to a specific XML format. For example, the instance `Dassault Systèmes` belonging to the entity type `Company`, has been added to its main concept name `Dassault Systemes`. The list representing the companies will be composed as such:

```
<concept name="~CompanyPart">
 <concept name="~CompanyPart.name">
  <concept name="~DassaultPart.name">Dassault Systemes</concept>
   <concept name="~DassaultPart.aliases">
       <concept_alias ID="1">DS</concept_alias>
        <concept_alias ID="2">Dassault Systèmes</concept_alias>
  </concept>
…
 </concept>
</concept>
```

Fig. 6. Companies' dictionnary updated

The XML lists will then be imported in the linguistic tool and each entity will be added to the linguistic resources as a new concept. The next time a document will be parsed to extract information, these entities would be recognized as they would be already recorded in the dictionaries of the information extractor. We infer that after a certain amount of time all the key information of the applied domain will be integrated in the knowledge base and the linguistic resources. The business application will converge towards a reliable knowledge base. The users will not have to validate as many information as in the beginning and thus would spend more time using the system for queries and publication. Consequently, the more the system will work, the more important the productivity gain will be.

5 Related Work

This research project presented in this document is innovating by many aspects and mainly by the ability to update the linguistic resources with the result of the knowledge base enrichment. Moreover, document annotation and knowledge base population are validated through the same interface, which provides an interesting ease-of-use and a consequent productivity gain for the user. Most systems are dealing with one aspect only, either the annotation issue or the knowledge enrichment one, instead of combining them.

Several methods have been proposed to extract the terminology of a domain or terms associations from texts and to use these extracted information to construct or enrich an ontology, such as OntoLearn [MIS 02]. But our system is not creating ontologies from scratch. On the contrary, considering a domain ontology, it can enrich the underlying knowledge base thanks to an IE tool. In OntoKnowledge [FEN 02] an RDF database stores the extracted information thanks to its IE tools whereas in our case the system uses the Topic Maps representation. Indeed, this language can better represent complex semantic relations when these have more than two roles such as a product adoption (company A uses the product B of the company C). Moreover, the knowledge base is constrained by the domain ontology.

The annotation of documents has been discussed around projects like Annotea[5] [KAH 01]. In that project, documents are annotated with comments and basic RDF metadatas such as author's name, date, source, etc. The user has to manually create its own annotations. But those systems are time-consuming and restricted about the kind of annotations. They cannot populate a knowledge base with domain pertinent information.

Commentaire [MSOffice2]: verifier le genre d'annotations de S-Cream et MnM

Other tools are using Human Languages Technologies elements like S-Cream [HAN 02] or MnM [VAR 02], Amilcare[6] [CIR 01] and Melita[7] [DIN 03], which assist the user when annotating textual documents from the Web. Those two systems are based on the Information Extraction algorithm (LP)^2 which uses machine learning to adapt the IE tool to new applications domains and to generalise induction rules. The algorithm needs an XML pre-tagged learning corpus for each of the user selected scenarios. The user corrects the new inserted tags until a precision threshold has been reached. They don't process the information extracted in the documents to populate a knowledge base but their algorithm provides a good solution for improving the information extractor's performance.

At the ISWC2003 conference, and more particularly in the workshop concerning "Human Language Technologies for the Semantic Web" (HLT4SW), several projects emerged around the problematic of using linguistic resources to create new Semantic Web applications. A common point between some of these projects is that they are based upon the GATE platform (General Architecture for Text Engineering) [CUN 02]. The main reason is that GATE provides open-source lexical, syntaxic and semantic resources to help constructing one's own linguistic tool. It is especially pertinent for the development of IE applications. That's why it is used in more and more Semantic Web projects that rapidly need an easy IE tool to do semantic annotation or ontology creation. But the annotation of semantic relations is not very performent yet and it can be an important lack for Semantic Web applications. Among the HLT4SW systems, KIM (Knowledge Information Manager) [POP 03] seems the nearest to our own approach. It extracts named entities from text, their attributes, aliases and some basic semantic relations such as the location of a person or an organization. Then it populates a knowledge base with this extracted information but does not annotate the documents with it. Another relative application is the Ar-

[5] Annotea Project website : http://www.w3.org/2001/Annotea
[6] Amilcare Project website : http://nlp.shef.ac.uk/amilcare
[7] Melita Project website : http://www.dcs.shef.ac.uk/~alexiei/Melita.htm

tequakt project [ALA 03] which searches the Web to answer a query on artists, extract the knowledge from all the pages found and after populating the knowledge base can generate the artist's biography thanks to a natural language generation tool. Compared to those two projects, our approach is more complete as we annotate documents and enrich the knowledge base at the same time. It is more powerful as we can extract complex semantic relations and precise information. But also because our system updates the linguistic resources with the new validated information in order to become more and more performant and reliable. It will also leverage the burden of annotating the documents from the users as the more the system will work the less they will spend time to annotate.

6 Conclusions and Future Work

We presented in this paper a new Semantic Web portal based on ontologies modelization that integrates a set of linguistic resources acting on the semantic annotation process. The benefits for the web communities will appear through queries and publications interfaces once the system will be well-trained on a specific domain. Indeed, through the ITM Semantic Web portal, our platform will provide semantic annotation, knowledge organization and management, visualization, publication and query services to non-expert users.

The use of linguistic resources can greatly facilitate the development of such applications and will become mandatory components for their success [BON 03]. For the moment, our solution only integrates the Information Extraction tool but we can think of other linguistic tools providing different interesting and pertinent services for the applications, such as the Categorization, the Summarization or the Natural Language Generation for the display of query results.

At last, our system is unique as the information extraction tool populates the knowledge base which in turn updates the linguistic resources with the new named entities extracted. As a business domain is mostly constrained by its actors, its vocabulary and its information, our system will permit to obtain a reliable knowledge base on this domain after a certain amount of time. Consequently, less human intervention during the validation step will be needed. It will result in improving the system productivity and in allowing the users to concentrate their efforts on other tasks, like exploiting the document annotations or the knowledge base.

The system is still under development and testing but we can already figure some future work on the update of the semantic relations and not only the named entities. Considering the ontological and terminological resources of the applied domain, the system will become able to construct its own linguistic resources and further on its extraction patterns [STE 03]. It will greatly facilitate the implementation of specific web portals of different web communities without being an expert in HLT.

7 References

[ALA 03]
ALANI H., KIM S., MILLARD D. and al., *Automatic Extraction of Knowledge from Web Documents*, in Proceedings of the Second International Semantic Web Conference, Workshop on Human Language Technology for the Semantic Web and Web Services, Florida, October 2003, pp. 77-88.

[BER 98]
BERNERS-LEE T., *Weaving the Web*, Eds HarperSanFrancisco, 1998, 226 p.

[BOU 03]
BOURIGAULT D, AUSSENAC-GILLES N. and CHARLET J, *Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas*, Revue d'Intelligence Artificielle, 2003, 24 p.

[BON 03]
BONTCHEVA K. and CUNNINGHAM H., *The Semantic Web: A New Opportunity and Challenge for Human Language Technology*, in Proceedings of the Second International Semantic Web Conference, Workshop on Human Language Technology for the Semantic Web and Web Services, Florida, October 2003, pp. 89-96.

[CIR 01]

CIRAVEGNA F., *Adaptive Information Extraction from Text by Rule Induction and Generalisation*, In Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001, 6 pp.

[CUN 02]

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. and al., *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, 2002, 8 pp.

[DIN 03]

DINGLI A., *Next Generation Annotation Interfaces for Adaptive Information Extraction*, In Proceedings of the 6th Annual Computer Linguists UK Colloquium (CLUK03), Edinburgh (UK), 6-7 January 2003, 5 pp.

[FEN 02]

FENSEL D., BUSSLER C., DING Y. and al., *Semantic Web Application Areas*, In Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, Stockholm, Sweden, 27-28 June 2002, 14 pp.

[GRI 01]

GRIVEL L., GUILLEMIN-LANNE S., LAUTIER C. and al., *La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux*, In Filtrage et résumé automatique de l'information sur les réseaux, 3ème congrès du Chapitre français de l'ISKO International Society for Knowledge Organization, 5-6 July 2001, 9 pp.

[HAN 02]

HANDSCHUH S., STAAB S. and CIRAVEGNA F., *S-CREAM – Semi-automatic CREAtion of Metadata*, In Proceedings of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002), ed Gomez-Perez, A., Springer Verlag, 2002, pp. 358-372.

[KAH 01]

KAHAN J., KOIVUNEN M., PRUD'HOMMEAUX E. and al., *Annotea: An Open RDF Infrastructure for Shared Web Annotations*, In Proceedings of the WWW10 International Conference, Hong Kong, May 2001, pp. 623-632.

[KAT 02]

KATZ B., LIN J. and QUAN D., *Natural Language Annotations for the Semantic Web*, In Proceedings of the ODBASE 2002, Irvine, California, October 2002, 15 pp.

[LAU 02]

LAUBLET P., REYNAUD C. and CHARLET J., *Sur Quelques Aspects du Web Sémantique*, Assises du GDR I3, Editions Cépadues, Nancy, December 2002, 20 pp.

[LU 02]

LU S., DONG M. and FOTOUHI F., *The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications*, Information Research, Special Issue on the Semantic Web, 7(4), 2002, 12 pp.

[MAE 03]

MAEDCHE A, STAAB S., STOJANOVIC N. and al., *SEmantic portal : The SEAL Approach*, In Spinning the semantic web : bringing the world wide web to its full potential, by D. FENSEL and al, The MIT Press, 2003, pp. 317-359.

[MIS 02]

MISSIKOF M., NAVIGLI R. and VELARDI P., *The Usable Ontology: an Environment for Building and Assessing a Domain Ontology*, In Proceedings of the First International Semantic Web Conference, Sardinia, Italy, June 2002, pp. 39-53.

[OWL 04]

JIM HENDLER J., HORROCKS I. and al., *OWL web ontology language reference*, W3C Recommendation 10 Feb 2004.

[POP 03]

POPOV B., KIRYAKOV A., MANOV D. and al., *Towards Semantic Web Information Extraction*, In Proceedings of the Second International Semantic Web Conference, Workshop on Human Language Technology for the Semantic Web and Web Services, Florida, October 2003, pp. 1-22.

[STE 03]

STEVENSON M. and CIRAVEGNA F., *Information Extraction as a Semantic Web Technology: Requirements and Promises*, In Proceedings of the 14th European Conference on Machine Learning (ECML-03), Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, 2003, 5 pp.

[VAR 02]

VARGAS-VERA M., MOTTA E., DOMINGUE J. and al., *MnM : Ontology Driven Tool for Semantic Markup*, In Proceedings of the Workshop Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002), Lyon (France), 22-23 July 2002.