Open access • Proceedings Article • DOI:10.1145/1088622.1088651

Document annotation and ontology population from linguistic extractions

Florence Amardeilh, Philippe Laublet, Jean-Luc Minel

Institutions: University of Paris, Centre national de la recherche scientifique

Published on: 02 Oct 2005 - International Conference on Knowledge Capture

Topics: Upper ontology, Ontology (information science), Suggested Upper Merged Ontology, Process ontology and Ontology-based data integration

Document annotation and ontology population from linguistic extractions

Florence Amardeilh, Philippe Laublet, Jean-Luc Minel

Document Annotation and Ontology Population from Linguistic Extractions

Florence Amardeilh
Mondeca[1], Laboratoire LaLICC[2],
R&D Lab[1], Paris, France.
florence.amardeilh@mondeca.com

Philippe Laublet
Laboratoire LaLICC[2],
Université Paris IV[2], France.
philippe.laublet@paris4.sorbonne.fr

Jean-Luc Minel
Laboratoire LaLICC[2],
CNRS, France.
jean-luc.minel@paris4.sorbonne.fr

ABSTRACT
In this paper, we present a workbench for semi-automatic ontology population from textual documents. It provides an environment for mapping the linguistic extractions with the domain ontology thanks to knowledge acquisition rules. Those rules are activated when a pertinent linguistic tag is reached. Those linguistic tags are then mapped to a concept, one of its attributes or even a semantic relation between several concepts. The rules instantiate these concepts, attributes and relations in the knowledge base constrained by the domain ontology. This paper deals with the underlying knowledge capture process and presents the first experiments realized on a real client application from the legal publishing domain.

Categories and Subject Descriptors
I.2.4 Knowledge Representation Formalisms and Methods – *Representation languages.*
I.2.7 Natural Language Processing – *Language parsing and understanding.*

General Terms
Languages, Experimentation, Theory.

Keywords
Knowledge acquisition tool, Knowledge extraction from text, Knowledge capture for the Semantic Web, Knowledge capture using natural language processing, Method for ontology population, Semantic Web.

INTRODUCTION
From the publishing industry to the competitive intelligence business, important volumes of data from various sources have to be processed daily and analyzed by professional users. First, they must select the most relevant documents with regards to their applications needs. Second, they must manually capture the pertinent knowledge contained in each selected resource. That knowledge is used to annotate the document by a set of descriptors (terms from the thesaurus like 'divorce' and named entities like 'Mr. Bouscharain') and to enrich the knowledge base (with named entities, attributes of these named entities and semantic relations between these named entities).

In the Semantic Web context, the content of a document can be described and annotated using knowledge representation languages such as RDF, XTM and OWL. RDF, the Resource Description Framework [15], is a formalism of knowledge representation from the semantic networks field. It is mainly used to describe resources, such as an electronic web document, by a set of metadata (author, date, source, etc.) and of descriptors. Those metadatas are composed of triples : (subject, verb, object) or (object 1, relation, object 2) or (resource, property, value) according to the needed description type.

Topic Maps are another formalism of knowledge representation [16]. Topic Maps define a set of topics linked to the same domain and constituting a semantic map of the knowledge. A topic represents everything that can be described or thought of by a human. It can participate in one or many relations, called associations, in which it plays a specific role. The topics have at least a name and intrinsic properties, called occurrences. This language allows a great flexibility in knowledge representation, especially with regards to modeling complex n-ary semantic relations.

OWL, Web Ontology Language [10], is used to formalize an ontology [8], or more generally some ontological and terminological resources [3], by defining concepts used to represent a domain of knowledge. Each concept is described by a set of properties, relations and constraints. The OWL formalism comes from some of description logic.

In our projects, we use RDF to describe the content of a resource, OWL to model the ontology which will represent an applicative or a functional vision of the domain and Topic Maps to implement the knowledge base that will contain the instances of the concepts, properties and relations described in the domain ontology. The pertinent knowledge of the domain, contained in the documents will be captured to instantiate the knowledge base and to create semantic annotations of these texts. The semantic annotations can then be interpretable by the machine to be later shared, published, queried or more generally used [13].

Semantic annotation and ontology population are greatly dependent of the knowledge captured in the documents by the professional users. Manual processing of documents is extremely expensive in time and resources. The entire process involves productivity and quality issues. For all those

reasons, companies are more and more looking for implementing solutions based on the use of linguistic tools that semi-automatically capture the pertinent information from textual documents.

Those natural language processing technologies should be tightly integrated into the future Semantic web applications and shall even become essential to the development, acceptance and use of the Semantic Web [2]. Thanks to the functionalities offered by the natural language processing technologies, and mainly those of Information Extraction, solutions adapted to the specific needs of Semantic Web might be developed such as:

- The semi-automatic construction of terminologies/vocabularies of a domain from a representative corpus as well as their maintenance [3].
- The semi-automatic enrichment of knowledge bases by named entities and semantic relationships extracted from textual documents [12].
- Semantic annotation of resources [11] [9] [17].

Contrary to the researches previously cited, we noticed in our own projects that the linguistic tools and the ontology of the client domain are independently modeled from one another. That's why we decided to implement a gateway between the concepts of the ontology and the semantic annotations produced by the linguistic tools that will capture the pertinent knowledge of the studied domain.

In this article, we present an innovative workbench for document annotation and knowledge acquisition. In the next section of this paper, we will describe the implementation of our solution. Then, we will present the results of the first experiments from a project in the legal publishing field. This project will be used throughout the paper to illustrate our work. Finally, we will sum up the results in order to develop a new hypothesis and conclude on the future perspectives of our research.

LINGUISTIC TOOLS INTEGRATION IN A WEB SEMANTIC PORTAL

The tools used: ITM™ & IDE™

Our solution is based on the Intelligent Topic Manager™ (ITM) tool from the company Mondeca. ITM™ is a software engineering platform for knowledge management. ITM™ integrates a semantic portal [1] providing four key functions : Editing, Search, Navigation and Publication. The domain ontology, formalized in OWL, constrains the knowledge base model, implemented in Topic Maps, the user interfaces as well as every functionality of the portal. The knowledge base elements point to their relative documents, accessible by URL on the Internet or in a content management system.

The linguistic analysis is done by the Insight Discoverer™ Extractor (IDE) developed by the company Temis. This tool implements a finite-state transducer method [7] that relies on a pre-treatment involving document segmentation in textual units (usually sentences), lemmatization and morpho-syntaxic analysis of those textual units. IDE™ produces a tagged conceptual tree (cf.Figure 3). Each node of that tree is named according to the semantic tag attributed to the textual unit extracted.

On the one hand, the ITM™ portal doesn't allow the (semi-)automatic enrichment of its knowledge base. On the other hand, the IDE™ information extractor, once the knowledge captured in a textual corpus, simply presents the extracted information to the user through an html interface without recording it in a knowledge base or even in a database for later reuse. Both companies then decided to collaborate on several projects (documentation, publishing, competitive intelligence, etc.). However, the customization of their tools for a client application is always done independently from one another, each having its own constraints.

Actually, Mondeca builds the domain ontology, if it doesn't already exist, according to the client needs and its existing data, whilst Temis develops specific linguistic resources for each application domain, reusing existing resources when possible (such as the named entities recognition tool). That's why the linguistic tags of the conceptual tree produced by IDE™ have different names from the concepts defined in the ontology even if they describe the same information. As a consequence, we must find a way to map one to the other in order to be able to instantiate the right concepts from the linguistic extractions.

ITM/IDE Integration in the Semantic Portal

The integration between the linguistic extractions from IDE™ and the ontological concepts of the domain defined in ITM™ must be achieved according to the following steps: 1) examining the conceptual tree resulting from the linguistic analysis; 2) defining the acquisition rules between linguistic tags and ontological concepts; 3) automatic processing of these rules on a corpus of documents (cf. Figure 1).

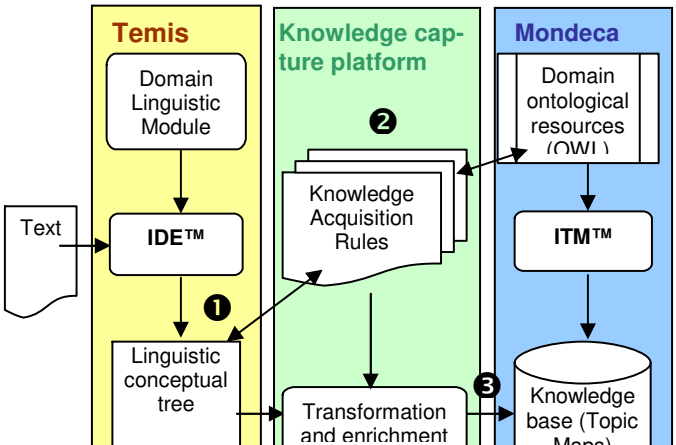

Figure 1. The process of Ontology population.

This process is applied to each of our client projects. We will illustrate this implementation through an example taken from one project about the French legal publishing domain: the author of legal articles must be aware of every legal text and court decisions. Thus, for every newly published document, a reference is recorded in the knowledge base with all its properties and linked to the other textual references cited.

Figure 2. Extract from a supreme court of appeal report.

The corpus used in our example is only composed of French legal decisions reports issued by supreme courts of appeal about divorces or employment contracts. The reports, cf. Figure 2, are divided in two parts: firstly a semi-structured header representing the information linked to this decision (date, supreme Court of appeal, decision number, appeal number, etc.) and then the unstructured document body describing the involved parties, the cause for complaint, the argumentation with the references to the legal coded texts (called « TC », e.g. « common Law ») and non coded texts (called « TNC », e.g. « Decree of the 30th of September 1953 »).

Conceptual tree from the linguistic analysis

As stated above, the IDE™ produces a conceptual tree from each linguistic analysis of a legal decision report (cf. Figure 3). Each node of this tree is semantically tagged and the textual value is included in parentheses. Our implemented solution parses this valued tagged tree in order to map the extracted information to an existing concept of the domain ontology, which can be a topic, an attribute, an association or a role in the knowledge base. To do this, we model the knowledge acquisition rules that will create an instance of an ontological concept at each corresponding node of the conceptual tree.

CIV. 1 D.S

COUR DE CASSATION

Audience publique du **23 mars 2004** Cassation partielle

M. BOUSCHARAIN, président Arrêt n° 510 F-D

Arrêt n° 510 F-D

Pourvoi n° F 02-19.839

(…) **REPUBLIQUE FRANCAISE**

AU NOM DU PEUPLE FRANCAIS

LA COUR DE CASSATION, PREMIÈRE CHAMBRE CIVILE, a rendu l'arrêt suivant :

Sur le pourvoi formé par Mme H, épouse Y, demeurant xxxxx, 75019 Paris, (…)

Sur le rapport de Mme G-L, conseiller référendaire, les observations de Me B H, avocat de Mme H, de la SCP V, avocat de la société P, les conclusions de Mme P, avocat général, et après en avoir délibéré conformément à la loi ;

Sur le moyen unique, pris en sa seconde branche :

Vu l'article L. 311-37 du Code de la consommation, dans sa rédaction antérieure à la loi n°2001-1168 du 11 décembre 2001 ; (…)

```
/REFERENCE DECISION(cassation 10400510)
   /FORMATION(CIV . 1)
      /Chambre civile(CIV . 1)
   /JURISDICTION(COUR DE CASSATION)
   /SESSION DATE(Audience publique du 23 mars
2004)
      /DATE(23 mars 2004)
         /MonthDayNumber(23)
         /month(mars)
         /YearNumber(2004)
   /Name lex(M. BOUSCHARAIN , président)
      Name(M. BOUSCHARAIN)
      role(président)
         /Role/Legal(président)
   /DECISION/SENTENCE(Arrêt n° 510 F-D)
         num(510 F-D)
   /APPEAL(Pourvoi n° F 02-19.839)
      num(F 02-19.839)
   …
   /REFERENCE(article L. 311-37 du Code de la
consommation)
      ref(article L. 311-37 du Code de la consom-
mation)
         /ARTICLE (article L. 311-37)
            art num(L. 311-37)
         TEXT(Code de la consommation)
            /CODE/Code consommation(Code de la
consommation)
```

Figure 3. Extract from a conceptual tree dealing with a legal decision.

The tree parsing is governed by some basic principles:

- A tree has necessarily a root, representing here the document or the main subject of the document (in our example, the decision itself).
- The tree parsing is a top-down parsing by a prefixed order : starting from the root, the algorithm parses first the left child before parsing the right child and so on recursively.
- Two parsings are necessary: the first one to capture the topics with their attributes and the second one to capture the associations with the roles played by the topics.

These two parsings are essential as not every topic necessarily plays a role in an association. Therefore, they wouldn't be instantiated if the parsing of the tree was only considering the associations, then their roles and finally the corresponding topics. In our example, it is especially the case of the topics « Person » having attributes such as « Name » and « Role » but not participating in any association as modeled in the client ontology.

Also we parse the topics with regards to their attributes and the associations wrt their roles in order to take advantage of the constraints modeled in the ontology. Indeed,

when finding a rule that instantiate a topic, we can automatically deduce the possible attributes of this topic from the ontology and search for the rules that might apply on these attributes. The same applies for the associations and their specific roles.

In order to process the conceptual tree, we choose, in a first step, to implement the knowledge acquisition rules in the Xpath language[1]. Indeed, this language allows us to parse a tree (XML document, conceptual tree, etc.), to directly reach any of its nodes and from any node to select any of its ancestors, descendants or siblings.

Definition of the Knowledge Acquisition Rules

Each node from the conceptual tree must be manually mapped with a concept of the domain ontology, whatever its type is (topic, attribute, association and role)[2]. To do this, we define the set of knowledge acquisition rules by hand. Those rules will set off the automatic creation of an instance of the ontological concept at each corresponding node of the conceptual tree. Table 1 sums up the various possible cases:

- A linguistic tag can be mapped into only one concept: « /art num » with the attribute « Num Article ».
- Many linguistic tags can be mapped into the same concept: « /Nom lex » and « /Noms lex » with the topic « Person ».
- A linguistic tag can be mapped into several concepts of the same type: « / COURT MEMBERS » with the topics « Legal person » and « Political person ».
- A linguistic tag can be mapped into several concepts of different types: « /REFERENCE » with the topics « Ref Editorial Legislative TNC » and « Ref Editorial Legislative TNC Article », with the association « Simple reference » and with the role « Targeted link ».
- A linguistic tag can't be mapped into any ontology concept: « /CAUSE COMPLAINT».
- A concept can't be mapped into a linguistic tag: the role « Originated Link ».

When a linguistic tag may instantiate more than one concept, the context of this tag, i.e. its parents, children or siblings nodes, helps resolve the ambiguities. For instance, if the node « /REFERENCE » has a child node « /ARTICLE », the topic « Ref Editorial Legislative TNC Article » will be instantiated, otherwise it will be the topic « Ref Editorial Legislative TNC ».

1 Site web du W3C : http://www.w3.org/TR/xpath

2 The vocabulary used here is the one of the Topic Maps.

Table 1. Examples of mappings between semantic tags and ontological concepts.

Name of the linguistic tag	Name of the concept in the ontology	Type in the kb	Context
/name lex	Person	Topic	
/names lex	Person	Topic	
/COURT MEMBERS	Legal person	Topic	If exists Descendant = /Legal
	Political person	Topic	If exists Descendant = /Political
/REFERENCE	Ref Editorial Legislative TNC	Topic	If not exists Child = /ARTICLE
	Ref Editorial Legislative TNC Article	Topic	If exists Child = /ARTICLE
	Simple Reference	Association	If exists Parent = /REFERENCE DECISION
	Targeted link	Role	If exists Parent = /REFERENCE DECISION
/art num	Num Article	Attribute	
/CAUSE COMPLAINT			
	Original link	Role	

The first part of a report, and therefore the linguistic extractions, deals with the supreme court of appeal decision. It contains all the attributes of the topic representing this decision, i.e. "Ref Editorial Case Law" marked by the tag « /REFERENCE DECISION ». It is then possible to map each of the nodes in this first part with the corresponding attributes, such as the tag « /FORMATION » with the attribute « formation » in Figure 4.

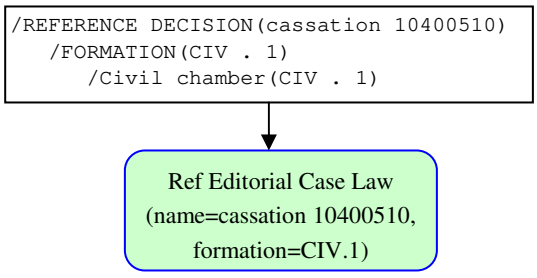

Figure 4. Mapping of a Topic from class "Ref Editorial Case Law" (translated).

The second part of the document deals with other types of concept instances, such as the persons (lawyers, presidents, counselors, etc.), and the references to the legal documents on which is based the argumentation of the dif-

ferent parties. Those references will be instantiated as a coded text or not, with their attributes (date, type of text, etc.), and then related to the decision through the association named « Simple Reference » and their role , i.e. « Targeted link », cf. Figure 5.

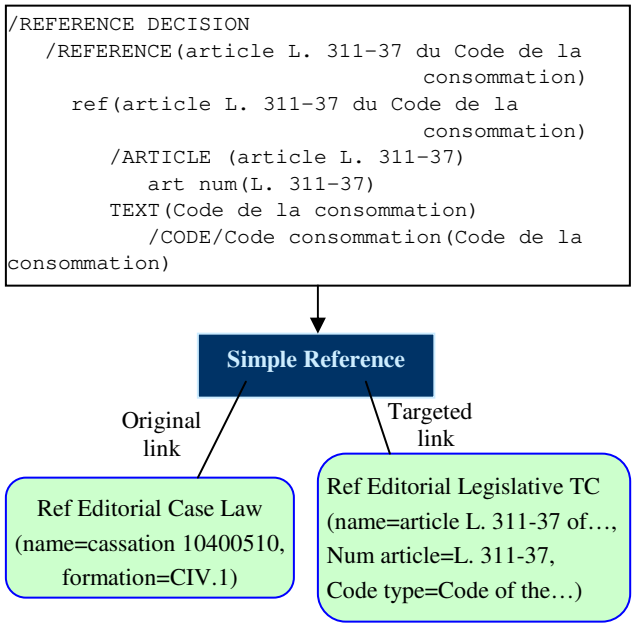

Figure 5. Mapping of a Association of type « Simple Reference » (translated).

Once the mapping is achieved, each acquisition rule will be formalized using Xpath and added in the domain ontology as a new attribute of the concept it will instantiate. For example, the concept "Ref Editorial Legislative TC Article" will have the following knowledge acquisition rule:

"/REFERENCE_DECISION/REFERENCE/ref[ARTICLE and TEXT]" which can be simplified in "//ref[ARTICLE and TEXT]".

Processing the Acquisition Rules

After processing the linguistic analysis on the documents, the resulting conceptual tree of each document is automatically parsed by the entire set of the knowledge acquisition rules. At each pertinent node, the corresponding instantiation of the knowledge base associated with every acquisition rule is processed. However, in order to avoid multi-creation of the same instance in the knowledge base, a check is done to verify its existence in the knowledge base. Once the tree parsing terminated, the user can visualize every new instance added to the knowledge base through a validation interface. From this interface, the user can modify and/or delete any created instance, and add another ones if missing. Thanks to this interface, the final user has the possibility to control the quality of the underlying knowledge base.

EXPERIMENTS AND RESULTS

Our experimental corpus is composed of 36 reports from supreme courts of appeal. On these 36 documents, given by the legal publishing company, 4 only were used to manually define 72 acquisition rules on 7 topic classes, 17 attributes types, one association type and two role types. There is an average of 3 acquisition rules per concept type. The other 32 documents were used as the test corpus. After reception of the linguistic module produced and compiled by the linguists from Temis, we processed the whole test corpus and obtain for each document its corresponding conceptual tree. We compared the linguistic tags with each instantiated concept to see which ones were correctly created, incorrectly created or even not created at all in the knowledge base. We want to point out the fact that we are not evaluating the linguistic extraction results or the quality of the ontology model but more specifically the performance of the knowledge acquisition rules themselves given this ontology model and this linguistic tool.

In order to evaluate quantitatively the results of this process, we used the precision and recall measures, previously defined to measure either information retrieval results (cf. TREC conferences), or information extraction results (cf. MUC conferences). In our case, we applied those measures to the tagged linguistic extractions with regards to the instantiated concepts in the knowledge base. Hence, we obtained the two following adapted measures:

- **Precision** measures the number of instances correctly acquired divided by the number of instances acquired.
- **Recall** measures the number of instances correctly acquired divided by the number of instances existing in the conceptual tree.

Following the analysis of the 32 documents from the test corpus, and with the same acquisition rules previously defined, the **Table 2** presents the results for the entire set of concepts found in the linguistic extraction corpus. A set of 1765 concepts of the ontology categorized in topics, attributes (or occurrences) of these topics, associations and roles have been detected in the conceptual trees of the test corpus. Among those concepts, 975 have been correctly instantiated, by the rules, 257 incorrectly instantiated and lastly 533 not instantiated. We are thus obtaining the following recall of 0,55 and a precision of 0,79.

Table 2. Experimentation results on the 32 documents of the test corpus.

Concept type	Number of concepts in the tree (A)	Number correctly instantiated (B)	Number incorrectly instantiated (C)	Number not instantiated (D)
Topics	585	432	139	14
Attributes	798	329	0	469
Associations	80	69	0	11
Roles	302	145	118	39
Total	**1765**	**975**	**257**	**533**

Table 3. Recall and precision measures.

Concept type	Recall (B/A)	Precision (B/B+C)
Topics	0.74	0.76
Attributes	0.41	1
Associations	0.93	1
Roles	0.48	0.55
Total	**0.55**	**0.79**

To sum up, even if the precision rate is satisfying for a first experimentation, we notice that an important number of textual units, correctly tagged in the conceptual tree, are not instantiated afterwards, especially the attributes and the roles. Other concepts, mainly topics, are also incorrectly instantiated. Despite technical problems concerning the parsing algorithm, incorrectness or missing instantiation are mainly due to the definition of the rules themselves. Indeed, part of the incorrectness is caused by a redundancy issue coming from conflicting rules. Another problem producing incorrect instantiations of the roles is the lack of control on the cardinalities modeled in the ontology inducing many roles of the same type instead of only one in the association.

We are also noticing that a lack of context awareness in the rules is responsible for not instantiating the different concepts, especially the topics and the attributes. It is essential to introduce more complexity in the acquisition rules based on the full context of the nodes in the generated conceptual tree. For the moment, our acquisition rules are limited to constraints on the child, parent or sibling nodes. Yet, the ancestor context is particularly important for the creation of topics' attributes. Let's take as an example the tag « /num » : if the direct parent node is « /ARTICLE », the attribute will be instantiated as an article number whereas if this same node is « /APPEAL », the attribute will correspond to an appeal number. The context of the descendant nodes can also bring more exactness with regards to the creation of a topic or an association. In Figure 6, the tag « /Names-of-persons » informs that the node deals with the class « Person » in the ontology. Therefore, this class has two sub-classes : « Legal person » and « Political person ». An analysis of the descendants of the node « /Names-of-persons », and mainly the presence of one or another nodes « Legal » or « Political », can set the right concept to instantiate.

```
/Names-of-persons(M. BOUSCHARAIN , president)
    Name(M. BOUSCHARAIN)
    role(president)
       /Role/Legal(president)
```

Figure 6. Contextual analysis example.

CONCLUSION AND DISCUSSION

This platform provides an innovative solution for ontology population from linguistic extractions thanks to the definition of acquisition rules. To our knowledge, there is no similar approach in the Semantic Web framework. Of course, other systems [12] are interested in ontology population thanks to linguistic tools but their ontologies are modeled according to the results of their linguistic extractions, at a higher level and without complex semantic relations (n-ary). At the contrary, our approach allows populating a given ontology (semi)-automatically taken any linguistic tool, once this one extracts the pertinent information about the domain as a conceptual tree (the IDE™ from Temis but also GATE[3]'s information extraction tool).

To sum up the process, the rule administrator compares all the ontology concepts and relations with all the possible paths of the conceptual tree resulting from the linguistic tool. He/She creates a first set of knowledge acquisition rules between the concepts that can be instantiated and the corresponding nodes. He/she tests it against a test corpus. He/She refines the set of knowledge acquisition rules according to the results obtained after testing. These two last steps are iterative until the acquisition rules, the ontology and the linguistic tool stabilize themselves in a version available for production. Then the whole system is delivered to the client and it becomes fully automatic, or semi-automatic if the client needs a validation step executed by the end-users.

The system needs a maintenance of the rules only when the linguistic resources and/or the ontology change. The rule administrator must be able to interpret a conceptual tree, to read an ontology and to construct Xpath rules. But it is not necessary for this person to be a specialist of the domain itself. To assist the maintenance process, the new instances added to the knowledge base can be automatically extracted to provide the linguistic tool with a list of instances by class of concepts so as to be added to its dictionaries. This is especially convenient concerning the named entities of the domain. For example, if 'Mr Bouscharain'

[3] General Architecture for Text Engineering, see http://gate.ac.uk/

has been added to the Knowledge base as a 'Person', the system sends this name with all the other new 'Person' to the linguistic tool that adds them in its 'Name lex' glossary. Hence next time that 'Mr Bouscharain' appears in a document, it will automatically be recognized and semantically tagged 'Name lex' in the conceptual tree.

Taken the issues raised during the first implementation of the system, we are defining the following priorities for our future research works:

- Improvement of the two conceptual tree parsings in order to manage more complexity in the rules thanks to a richer contextualisation.
- Detection of the conflicts caused by recovery problems between rules.
- Checking the respect of the cardinalities, especially for the roles of an association.

The achievement of some of these priorities would rapidly improve the actual system performance, mainly with regards to associations and roles. There is still the problem of coherence and maintenance between the knowledge acquisition rules that might become more and more numerous according to the size of the domain ontology to populate. The manual definition of all the acquisition rules is itself heavy and error-prone. And if the linguistic resources or if the client ontology are modified, then all those rules must be verified and updated by the administrator of these rules.

To move the system to a complete different domain, the same method as defined above need to be applied. The concepts of the ontology and the semantic tags of the conceptual tree might be completely different as they are dependant of this new domain. So the rule administrator must define a completely new set of knowledge acquisition rules based on these new ontology and new conceptual tree. We also tested the system on some wider domains such as competitive intelligence or professional press publishing. Even if there are more concepts to instantiate and a richer conceptual tree, we noticed that it was not an exponential work to define the set of acquisition rules for those domains.

That's why we propose to develop a formal language to describe the knowledge needed to populate an ontology from a conceptual tree. This language will be inspired from LangText [4], developed to model the linguistic knowledge in the Contextual Exploration framework [5], [14]. One of the advantage of this language is the declarative way to formalize the notions of search space, of indicator and of annotation of a textual unit (word, phrase, sentence, paragraph,…).

Actually, it is necessary to adapt this language to a conceptual tree parsing and not a free text document. This adapted language will allow a better knowledge maintenance, a greater efficiency in the definition of the concepts to instantiate by the potential conflicts management, and thus a productivity gain for the user. This language is still under development at this time, but the Figure 7 presents a knowledge acquisition rule with the actual primitives that are necessary to instantiate a concept of the ontology in the corresponding knowledge base.

```
RuleName = R6 ;
NodeIndicator = REFERENCE
OntologyConceptType : Topic ;
OntologyConceptClass = Ref Editoriale
                       Legislative TNC
ContextDependency :
{NotExist :
        [SpaceSearch : child]
        [clueNode : ref]
}
{Exist :
        [SpaceSearch : descendant]
        [clueNode : ARTICLE]
}
endRule
```

Figure 7. Knowledge Acquisition Rule Formalism.

Lastly, we would like to emphasize the fact that this system must stay generic enough so as to be able to define and apply the acquisition rules to any application domain. The purpose of these knowledge acquisition rules is to transform a linguistic tag into an instantiated concept of the domain ontology.

REFERENCES

[1] Amardeilh F., Francart T, *A Semantic Web Portal with HLT Capabilities*, In Actes du colloque « Veille Stratégique Scientifique et Technologique » (VSST2004), Toulouse, France, Vol. 2, p 481-492, (2004).

[2] Bontcheva K., 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, 20-23 October 2003, p. 89-96, (2003).

[3] Bourrigault D., Aussenac-Gilles N., Charlet J., *Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas*, In Revue d'Intelligence Artificielle, 18(4), 24 pp, (2004).

[4] Crispino G. "Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes", PhD Thesis, Paris 4-Sorbonne University, 241pp, (2003).

[5] Desclès J.-P., Jouis C., Oh H-G. et al., *Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte*, Knowledge modeling and expertise transfer, Amsterdam, p. 371-400, (1991).

[6] Grisham R., Sundheim B., *Message understanding conference - 6: A brief history*, In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, p.466-471, (1996).

[7] Grivel L., Guillemin-Lanne S., Lautier C. et 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, Paris, 5-6 July 2001, 9 pp, (2001).

[8] Gruber T., *A Translation approach to portable ontology specifications*. In Knowledge Acquisition, 5(2), p. 199-220, (1995).

[9] Handschuh S., Staab S., Ciravegna F., *S-CREAM – Semi-automatic CREAtion of Metadata*, In Proceedings of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Spain, 1-4 October 2002, Springer Verlag, p. 358-372, (2002).

[10] Hendler J., Horrocks I. and al., *OWL web ontology language reference*, W3C Recommendation, (2004), available at <http://www.w3.org/TR/owl-ref/>

[11] Kahan J., Koivunen M., Prud'Hommeaux E. et al., Annotea: An Open RDF Infrastructure for Shared Web Annotations, In *Proceedings of the WWW10 International Conference*, Hong Kong, May 2001, p. 623-632, (2001).

[12] Kiryakov A., Popov B., Ognyanoff D., et al., *Semantic Annotation, Indexing, and Retrieval*, In Proceedings of the 2nd International Semantic Web Conference (ISWC2003), Florida, 20-23 October 2003, p.484-499, (2003).

[13] Laublet P., Reynaud C. et Charlet J., *Sur quelques aspects du Web Sémantique*, In Assises du GDR I3, Eds Cépadues, Nancy, 20 pp, (2002).

[14] Minel J.-L., J-P. Desclés, E. Cartier, et al., *Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText*, In Revue Technique et Science informatiques, 20(3), Hermès, p. 369-395, (2001).

[15] Ora L., Swick R., *Resource Description Framework (RDF) Model and Syntax Specification*, W3C Recommendation (1999), available at <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>

[16] Park J., Hunting S., *XTM Topic Maps : Creating and using Topic Maps for the Web*, Addison Wesley, Boston, p. 81-101, (2003).

[17] Vargas-Vera M., Motta E., Domingue J., *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, p. 43-47, (2002).