TtoO: Mining a thesaurus and texts to build and update a domain ontology

Josiane Mothe,
Institut de Recherche en Informatique de Toulouse.
118 route de Narbonne
31062 Toulouse Cedex 9, France
Phone : +33561556444
Fax : +33561556258
mothe@irit.fr

Nathalie Hernandez
Institut de Recherche en Informatique de Toulouse.
118 route de Narbonne
31062 Toulouse Cedex 9, France
Phone : +33561556899
Fax : +33561556258
hernandez@irit.fr

TtoO: Mining a thesaurus and texts to build and update a domain ontology

Josiane Mothe, Nathalie Hernandez
Institut de Recherche en Informatique de Toulouse.

ABSTRACT : This chapter introduces a method re-using a thesaurus built for a given domain, in order to create new resources of a higher semantic level in the form of an ontology. Considering ontologies for data-mining tasks relies on the intuition that the meaning of textual information depends on the conceptual relations between the objects to which they refer rather than on the linguistic and statistical relations of their content. To put forward such advanced mechanisms, the first step is to build the ontologies. The originality of the method is that it is based both on the knowledge extracted from a thesaurus and on the knowledge semi-automatically extracted from a textual corpus. The whole process is semi-automated and experts' tasks are limited to validating certain steps. In parallel, we have developed mechanisms based on the obtained ontology to accomplish a science monitoring task. An example will be given.

KEYWORDS: Knowledge Discovery, Data Exploration, Data Mining, Ontology

INTRODUCTION

Scientific and technical data, whatever their types (textual, factual, formal or non-formal), constitute strategic mines of information for decision makers in economic intelligence and science monitoring activities, and for researchers and engineers (e.g. for scientific and technological watch). However, in front of the growing mass of information, these activities require increasingly powerful systems offering greater possibilities for exploring and representing the collected information or extracted knowledge. Upstream, they must ensure search, selection and filtering of the electronic information available. Downstream, when communicating and restituting results, they must privilege ergonomics in presentation, exploration, navigation and synthesis.

To be manipulated by advanced processes, textual document content has first to be represented synthetically. This process is known as indexing and consists in defining a set of terms or descriptors that best correspond to the text content. Descriptors can either be extracted from the document themselves or by considering external resources such as thesauri. In the first approach, which can be fully automatic and thus more appropriate for large volumes of electronic documents, the texts are analyzed and the most representative terms are extracted (Salton, 1971). These are the terms which constitute the indexing language. The automatic weighting of index terms (Robertson and Sparck-Jones, 1976), their stemming (Porter, 1981), the automatic query reformulation by relevance feedback (Harman, 1992) or per addition of co-occurring terms (Qiu and Frei, 1993) in information retrieval systems (IRS) are methods associated with automatic indexing and have been effective as attested by international evaluation campaigns such as Text Retrieval Conference (trec.nist.gov). One common point of all these approaches is that they make the assumption that the documents contain all the knowledge necessary for their indexing. On the other hand, thesauri are used to control the terminology representing the documents by translating the natural language of the documents into a stricter language (documentary language) (Chaumier, 1988). A thesaurus is

based on a hierarchical structuring of one or more domains of knowledge in which terms of one or more natural languages are organized by relations between them, represented with conventional signs (AFNOR 1987). With ISO 2788 and ANSI Z39, their contents can be standardized in terms of equivalence, hierarchical and associative relations between lexemes. Indexing based on a thesaurus is generally carried out manually by librarians who, starting from their expertise, choose the terms of the thesaurus constituting the index of each document read. In an IRS, the same thesaurus is then used to restrict the range of a query or, on the contrary, to extend it, according to the needs of the user and the contents of the collection. Other types of systems combine the use of a thesaurus with classification and navigation mechanisms. The Cat a Cone system (Hearst, 1997) or IRAIA system (Englmeier and Mothe, 2003) make use of the hierarchical structure of the thesaurus to allow the user to browse within this structure and thus to access the documents associated with the terms. Compared to automatic indexing, the thesaurus approach leads to more semantic indexing, as terms are considered within their context (meaning and related terms). However, using thesauri raises several problems: they are created manually and their construction requires considerable effort; their updating is necessary; their format is not standardized: ASCII files, HTML, data bases coexist; finally, thesauri have a weak degree of formalization since they are built to be used by domain experts and not by automatic processes. Various solutions to these issues have been proposed in the literature. Automatic thesauri construction can call upon techniques based on term correlation (Tudhope, 2001), document classification (Crouch and Yang, 1992) or natural language processing (Grefenstette, 1992). On the other hand, the standards under development within the framework of the W3C as SKOS Core (Miles et al., 2005) aim at making the thesaurus migrate towards resources that are more homogeneous and represented in a formal way by using OWL language (McGuiness and Harmelen, 2004) and making these resources available on the Semantic Web. However, thesauri represent a domain in terms of indexing categories and not in terms or meaning. They do not have a level of conceptual abstraction (Soergel et al., 2004) which plays a crucial role in man-machine communication. Ontologies make it possible to reconsider this problem since it is a “formal and explicit specification of a shared conceptualization” (Fensel, 1998). Automatic semantic indexing, starting from concepts rather than terms that are frequently ambiguous, then becomes possible (Aussenac and Mothe, 2004). However, the development of an ontology is costly as it requires many manual interventions. Indeed, ontology construction techniques in the literature generally do not base the development of ontologies on existing terminological representations of the field but on a reference corpus which is analyzed (Aussenac et al., 2000).

In this chapter, we propose to present the method we have developed and implemented to re-use a thesaurus built for a given domain, in order to create new resources of a higher semantic level in the form of a domain ontology. The originality of the method is that it is based both on the knowledge extracted from a thesaurus and on the knowledge automatically extracted from a textual corpus. It thus takes advantage of the domain terms stated in the thesaurus. This method includes the incremental population and update of an ontology, by the mining of a new set of documents. The whole process is semi-automated and experts’ tasks are limited to validating certain steps. In parallel, we have developed mechanisms based on the obtained ontology to accomplish a science monitoring task. We will give an example in order to illustrate the added value of using ontologies for data mining. This chapter is organized as follows: section 2 presents the main differences between thesauri and ontologies. In section 3, we review the literature related to ontology building and ontology integration for representing and accessing documents. Section 4 presents our method for building the ontology from the thesaurus. Section 5 explains how we propose to mine documents to update the ontology.

Finally, in section 6, we present our case study: building a lightweight ontology in astronomy by transforming the IAU thesaurus and developing a prototype enabling a science monitoring task on the domain.

THESAURI AND ONTOLOGIES

The main distinction between a thesaurus and an ontology is their degree of semantic engagement (Lassila and McGuiness, 2001). This degree corresponds to the level of formal specification restricting the interpretation of each concept, thus specifying their semantics.

Standardization of thesauri content

The ISO 2788 and ANSI Z39 (http://www.techstreet.com/cgi-bin/pdf/free/228866/z39-19a.pdf) standards have proposed the guiding principles for building a thesaurus. It is a terminological resource in which terms are organized according to restricted relations (Foskett, 1980): equivalence ($term_i$ "*use for*" *referring* to $term_j$, *instead of* $term_j$ "*use*" $term_i$), hierarchical term1 ($term_k$ "*broader term*" *of* $term_l$, $term_l$ "*narrower term*" of $term_k$), and cross reference ($term_m$ "*related term*" *to* $term_n$). The relations present in a thesaurus meeting those standards are shown in figure 1.

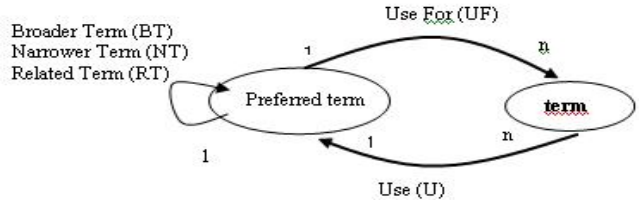

Figure 1 Relations between terms in a thesaurus

From the point of view of knowledge representation, thesauri have a low degree of formalization. The distinction between a unit of meaning (or concept) and its lexicon is not clearly established. Synonymy relations are defined between terms but concepts are not identified. This can be explained by the initial use of thesauri which was to express how a domain can be understood not in terms of meaning but in terms of terminology and categories to help the manual indexing of domain documents. Moreover, the semantics associated to a thesaurus are limited. Relations are vague and ambiguous. The semantic links between terms often reflect the planned use of the terms rather than their correct semantic relations. The relation "*broader term*" can, for example, include the relations « *is an instance of* », « *is a part of* » (Fischer, 1998). The associative relation «*related term*» is often difficult to exploit as it can connect terms suggesting different semantics (Tudhope et al., 2001). For example, in the BIT thesaurus (http://www.ilo.org/public/libdoc/ILO-Thesaurus/french/tr1740.htm) describing the world of work, the term « family » is related to the term « woman » and « leave of absence », the semantic relation between those pairs of terms is intuitively different. Because of the choices made during their elaboration, thesauri lack both formalization and coherence compared to ontologies.

Ontologies

Instead of representing a domain in terms of indexing categories, the aim of ontologies is to propose a sound base for communication between machines but also between machines and humans. They consensually define the meaning of objects firstly through symbols (words or phrases) that designate and characterize them, and secondly through a structured or formal representation of their role in the domain (Aussenac 2004). Before presenting ontologies

through their degree of formalization (lightweight and heavyweight ontologies), we define the three units of knowledge on which they rely: concept, relations and axioms.

Concepts

A **concept** represents a material object, a notion or an idea (Uschold and King, 1995). It is composed of three parts: one or several **terms**, a notion and a set of objects. The notion corresponds to the concept's semantics and is defined through its relations to other concepts and its attributes. The set of objects corresponds to the objects defined by the concept and is called concept extension; objects are concept instances. The terms designate the concepts. They are also called concept **labels**. For example, the term « chair » references an artifact, piece of furniture on which one can sit and all objects having this definition. In order to identify a concept with no ambiguity, it is recommended to reference a concept with several terms that eliminate the problem of synonymy and to disambiguate the meanings of terms by comparing them to each other (Gomez-Perez et al., 1996).

Relations between concepts

A semantic relation R represents a kind of interaction between domain concepts c1, c2,... cn.
The notion of *subsumption* (also called "*is a*" or taxonomic relation) is a binary relation implying the following semantic engagement: a concept c1 "s*ubsumes"* a concept c2 if all the semantic relations defined for c1 are also semantic relations of c2, in other terms if concept c2 is more specific than concept c1.
« *Associative* » relations are interaction relations between two concepts that are not subsumption relations. This corresponds to the notion of role in description logics and makes it possible to characterize concepts. These relations are either relations between concepts or relations between a concept and a data type. The semantics associated to it are referenced by a label. Examples of associative relations are "*is a part of*", "*is composed of*", "*has a color*"… They can also be specified with logic properties associated to the relation such as transitivity.

Axioms

Axioms aim at representing in a logic language the description of concepts and their relations. They represent the intention of concepts and relations and the knowledge that does not have a terminological character. Their specification in an ontology can have different purposes: define the signature of a relation (domain and range of the relation), define restriction on values of a relation, define the logic property of a relation (transitivity, symmetry…). Through axioms both the consistency of the knowledge stated in the ontology is checked and new knowledge inferred.

Lightweight ontologies vs Heavyweight ontologies

Lightweight ontologies are composed of two semiotic levels (Maedche and Staab, 2002). The lexical level (L) covers all the terms or labels defined to designate the concepts, enabling their detection within textual contents. The conceptual level defined in the structure (S) of the ontology represents the concepts and semantics defined from the conceptual relations that link them.
The structure of an ontology is a tuple $S := \{C, R, \leq_C, \sigma_R\}$ where :

- C, R are disjoint sets containing concepts and associative relations
- $\leq_C$: C x C is a partial order on C, it defines the hierarchy of concepts

  $\leq_C$ (c1, c2) meaning that c2 "is a" c1, it is called a taxonomic relation
- σ_R : $R \rightarrow C \times C$ is the signature of an associative (or non-taxonomic) relation

The lexicon of a lightweight ontology is a tuple $L : \{L^C, L^R, F, G\}$

- L^C, L^R are disjoint sets containing labels (or terms) referencing concepts and relations
- F, G are two relations called reference, they enable access to the concepts and relations designated by a term and vice versa.
- $F \rightarrow L^C$ for the concepts and $G \rightarrow L^R$ for the relations
  - For $l \in L^C$, $F(l)=\{c / c \in C\}$
  - For $c \in C$, $F^{-1}(c)=\{l / l \in L^C\}$
  - For $l \in L^R$, $G(l)=\{r/r \in R\}$
  - For $r \in R$, $G\text{-}1(r)=\{l / l \in L^R\}$

Heavyweight ontologies are based on the same structure and lexicon as lightweight ontologies and also include a set of axioms. This kind of ontology considerably restricts the interpretation of concepts, thus limiting ambiguity. However, their construction is extremely time-consuming and thus they cover only precisely defined domains. For data representation and data mining, considering lightweight ontologies is the first step in integrating formal knowledge in systems. Lightweight ontologies can play the role formerly played by thesauri by defining an indexing language relying on a more formalized knowledge that can be exploited in the retrieval or mining process. Moreover, as described in the next section, lightweight ontologies are easier to construct.

RELATED WORK : ELABORATING ONTOLOGIES

Designing ontologies is a difficult task involving the development of elaborate processes that extract domain knowledge that can be manipulated by information systems and interpreted by humans. Two kinds of design exist: entirely manual design and design based on learning stages.

Several assumptions and methodologies have been defined to facilitate manual design. The assumptions rely on philosophical characteristics and follow collaborative modeling processes (Guarino et al., 1994). However, this approach is time consuming and causes maintenance and update problems (Ding, 2002). Over the last decade, ontology learning has emerged as a sub domain of knowledge engineering. Thanks to technological progress in the domain of Information Retrieval, automatic learning and natural language processing, attempts are now being made to make the elaboration of ontologies semi-automatic. Ontology learning generally leads to the development of lightweight ontologies. In (Maedche and Staab, 2001), different types of approaches are distinguished in function of the resources used to support the elaboration: texts, dictionaries or thesauri, knowledge bases, semi-structured schemas and relational schemas. We have focused on approaches based on texts and thesauri as they offer, in our point of view, the possibility of taking advantage of existing resources and up-dating ontologies.

In order to transform a thesaurus into an ontology, existing approaches aim at capturing the informal semantics of the thesaurus manually (Wielinga et al. 2001), with syntactic patterns (Soergel et al. 2004), or with inferences (Hahn and Schulz, 2004). Considering the time it requires, an entire manual process is conceivable only in specific cases. A semi-automatic treatment appears more adapted. A first contribution by (Soergel et al. 2004) aims at facilitating the transformation. However, the manual work required of the experts remains considerable as they have to analyze each pair of terms in order to extract patterns that explicit their semantic relations. Our contribution limits the manual work by defining syntactic patterns for generic concepts instead of for each pair. Moreover, a drawback of the literature approaches is that the ontology is elaborated only according to the knowledge stated in the thesaurus. This knowledge does not necessarily reflect the evolution of knowledge in

the domain considered. We thus propose to transform a thesaurus from the knowledge it contains but also from information stated in a reference document corpus of the domain.
Concerning ontology elaboration from texts, different tools have been developed. Each presents different functionalities and has enabled the elaboration of various ontologies. Text-To-Onto, developed by the AIFB Institute of Karlsruhe University, is an application based on the extraction of knowledge from web documents that considers the reuse of existing ontologies (Maedche and Staab, 2001). It is integrated in the KAON platform that performs the editing and maintenance of ontologies (Bozsak et al., 2002). OntoBuilder (Gal et al., 2004), developed by the Technion of Haifa, helps to construct ontologies from XML files guided by a refining stage done by the user.
The TERMINAE methodology (Aussenac et al., 2000) proposes an approach for selecting concepts and their relations in texts. It uses natural language processing tools to detect relevant terms from documents and their syntactic relations. Terms are then clustered according to their syntactic context thus helping in the identification of concepts and relations between them. As presented in section 4, our methodology extends TERMINAE by integrating terminological resources such as thesauri.

FROM A THESAURUS TO AN ONTOLOGY

General Framework

In order to specify the key points in the process of transforming a thesaurus into an ontology, we present our approach in the framework of the methodology TERMINAE (Aussenac et al., 2000). This methodology that is well known for ontology elaboration from texts involves five steps.
The aim of the *first step* is to specify the needs the ontology must satisfy. In the case of the transformation of a thesaurus into a lightweight domain ontology to represent and access documents, the needs identified are :

- identifying domain terms and their lexical variants in order to detect them in documents,
- grouping these terms as concepts in order to determine the objects and notions referenced in the documents,
- structuring the concepts through taxonomic and associative relations in order to guarantee a good semantic indexation,
- formalizing the ontology in a language understood by the systems so that they will be able to manipulate it.

The *second step* lies in the choice of reference corpus from which the ontology will be built. This choice is a deciding factor in the construction of ontologies. The corpus must describe the items of knowledge that will be integrated into the ontology. When a thesaurus is being transformed, the corpus must satisfy two conditions. Firstly, it must be possible to capture the implicit knowledge that is not formalized in the thesaurus. Secondly, the corpus must help the updating process through recent documents in the domain. In our approach, the corpus is extracted from existing corpuses and experts must guarantee that it covers the whole domain in a representative period. Abstracts of articles published in journals of the domain describe this type of information. Full articles may be used but the advantage of abstracts is that the information they contain is condensed.
The *third step* is that of the linguistic study of the corpus. The aim is to extract from the documents the representative domain terms and their relations (lexical and syntactic) using appropriate tools. At the end of this step, a set of terms, the relations between the terms and clusters are obtained. For the transformation of a thesaurus, this step integrates the knowledge represented in the thesaurus as the terms it contains are representative of the domain and may be grouped according to their relations. The linguistic study of the reference corpus is also

necessary for the extraction of terms which are not present in the thesaurus and the relations between terms which are not explicited there. For this analysis, we used the syntactic analyzer Syntex (Bourigault and Fabre, 2000). It presents the advantage of being based on endogenous learning to carry out analyses on different domains. It extracts phrases from the documents and the context of their occurrence (the words they dominate and by which they are dominated). A method must be elaborated, however, to define the mechanisms to select the terms and their relations from the knowledge extracted from the thesaurus and the information extracted from the corpus. This is done by the method we propose.
The *fourth step* corresponds to the normalization of the results obtained by the previous step. Concepts and semantic relations are defined from the terms and lexical relations. At this stage, the thesaurus can be used to specify the concepts.
The *final step* concerns formalization: the semantic network defined in the previous step is translated into a formal language and in this case can be done using OWL. The advantage of this language is that it is composed of three sub-languages of an incremental level of formalization. The use of OWL-Lite gives a first formalization of the ontology which can evolve.

To transform a thesaurus, the method we propose implements steps 3, 4 and 5 of the methodology. Our implementation relies on three hypotheses based on the re-use of thesaurus relations:

- the preferred terms are the main terms in the domain and are clues for the constitution of terms designating domain concepts,
- the relations between terms and preferred terms are relations of synonymy between terms, they enable the grouping of terms as being the label of a particular concept,
- the relations between preferred terms are clues for the definition of relations between concepts.

Figure 2 schematizes the overall process. It is based on a mechanism which can be decomposed into several phases described in the following section.

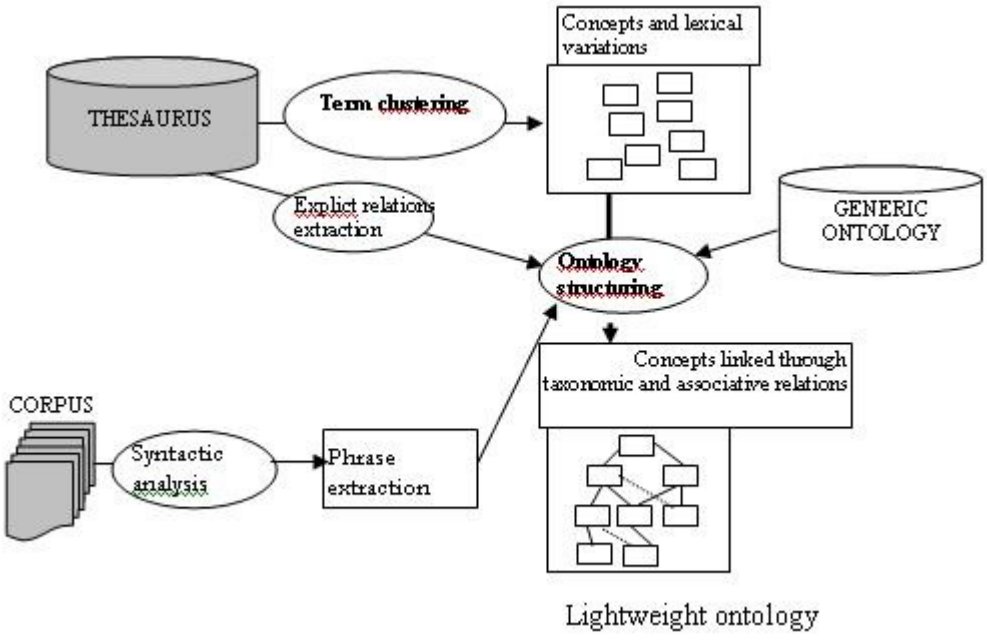

Figure 2. Overall process

Defining Concepts And Their Labels

Concepts and concept labels are key elements of an ontology. This stage aims to extract from the thesaurus lexicon a conceptualization in order to define a first set of the ontology concepts. In order to do so, terms defined as "preferred terms" and the relations "use for" and "use" are analyzed. We consider these relations as synonym relations.

Term clusters are made from each preferred term and the set of terms they are linked to according to the relations UF and U. Rule R1 represents this process.

If t3 U t1 **then** t1 and t3 are grouped, with t1 preferred term.
If t1 UF t2 **then** t1 and t2 are grouped, with t1 preferred term
(R1)

The latter are then aggregated according to the transitive closure of the U and UF relations. If a preferred term has led to a first cluster and appears in another cluster, both clusters are grouped. The transitive closure consists in grouping terms by rule R2.

If t1 UF t2 and t2 UF t3, **then** t1 UF t3 **=>** t1, t2 and t3 are grouped,
with t1 as principal preferred term
If t4 U t5 and t5 U t6 **then** t4 UP t6 **=>** t4, t5 and t6 are grouped,
with t6 as principal preferred term
(R2)

Structuring These Concepts: Taxonomic Relations Extraction

Some taxonomic relations are directly extracted from the explicit relations of the thesaurus. A higher hierarchical level is added thanks to the analysis of syntactic labels and the creation of generic concepts. Figure 3 schematizes these mechanisms.

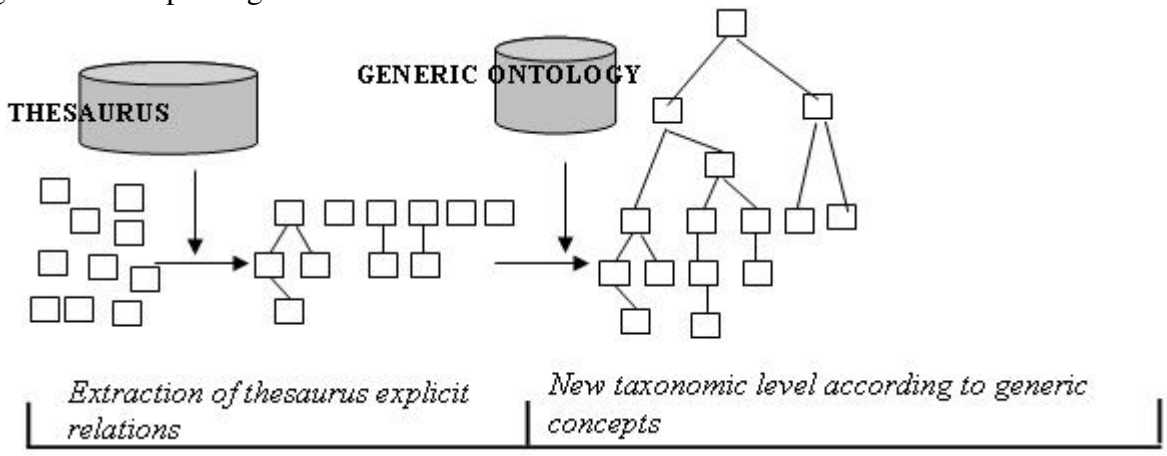

Figure 3 Taxonomic relations extraction process

Extracting Explicit Taxonomic Relations

In order to extract taxonomic relations from the thesaurus, thesaurus « Broader Term » and « Narrower Term » relations are considered. These relations defined between terms, now concept labels, are used to define candidate taxonomic relations between the concepts whose labels are linked in the thesaurus (see R3). These relations have to be analyzed carefully as they can imply the « is part of » and « is an instance of » relations. Our work does not deal with the automatic disambiguation of these relations. Note that several domain thesauri consider the « broader term » and « narrower term » in their strict meaning.

If t1 « narrower term » t2 with t1 label of concept c1 and t2 label of concept c2
then c2 « is a » c1
If t3 « broader term » t4 with t3 label of concept c3 and t4 label of concept c4
then c3 « is a » c4
(R3)

New Taxonomic Level: Generic Concepts

A drawback of thesauri is that their highest hierarchical level generally contains a huge number of terms (Soergel et al., 2004). These terms are the ones for which no « broader term » relations have been defined. This is explained by the fact that thesauri do not define generic categories enabling the classification of the domain terms. This drawback is also noticed in ontologies obtained by the transformation of thesauri, causing problems when the user chooses to explore the ontology from top to bottom. The first hierarchical level is vast, making the start of the browsing difficult. For example, the highest level of the transformation of the IAU thesaurus used to evaluate our approach contains more than 1100 concepts.

We thus propose to add a new hierarchical level to ease the browsing of the ontology. We propose the definition of generic concepts which characterize top level concepts. A generic concept refers to a concept of the domain or a concept added to structure the ontology. In (Soergel et al., 2004), generic concepts are defined according to a categorization schema existing in the domain. This process cannot be applied to all domains, as they do not always exist. Moreover, it implies manual work by the expert who has to map the ontology concepts to the created generic concepts. We propose a new, more automatic approach. A generic ontology (such as WordNet (Miller 1988) or DOLCE (Gangemi et al., 2002), is used to define generic concepts. First, the level 0 concepts of the built ontology are mapped to the concepts of the generic ontology. The generic concepts are then defined according to the most specific ancestors of those concepts in the generic ontology. Figure 4 schematizes the process.

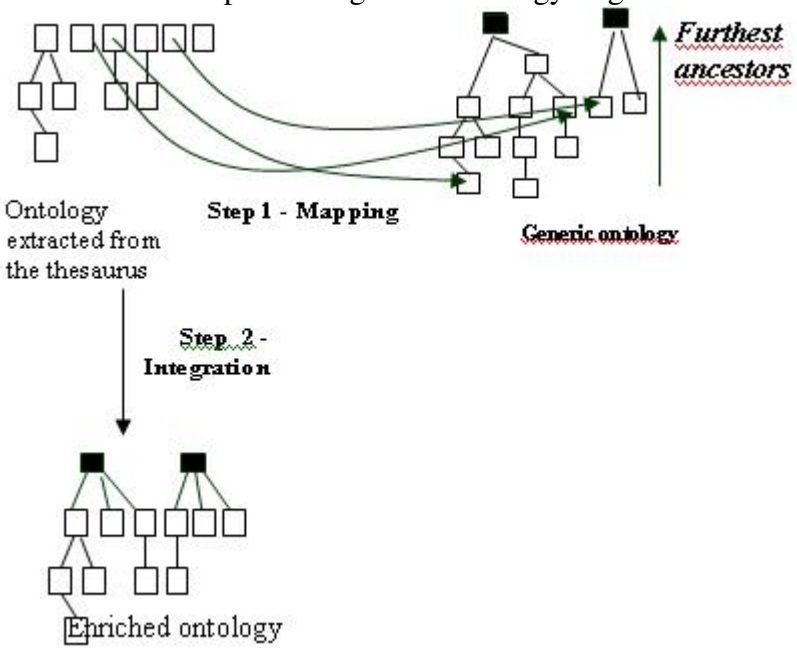

Figure 4 Adding an abstraction level to the ontology

Formula R4 synthesizes the entire process.

If c ⇔ sw with sw∈ { WordNet_concepts}
then c « is a sub-concept of » ta
with ta the most specific hypernomym of sw
(R4)

Generic Concepts Definition
We have developed a method to map the ontology concepts to concepts of WordNet. This resource is a lexical reference system whose construction was inspired by linguistic theories. It covers the general English language domain.
The labels of the top concepts in the ontology are used to select candidate concepts in WordNet. A disambiguation process is defined to disambiguate labels that refer to different WordNet concepts. This process relies on:

- the glossary provided by WordNet to describe in natural language the meaning of the concepts
- the descendant concepts of each concept defined by the WordNet hyperonymy relation
- the ancestor concepts of each concept defined by the WordNet hyponymy relation.

General terms referring to the considered domain treated by the ontology are first specified by experts. They are then looked for in the glossary associated to each candidate concept.
The following three propositions are successively applied:

- Proposition 1: If one of the terms is found in its glossary, the candidate concept is automatically chosen for the mapping. Otherwise, proposition 2 is applied.
- Proposition 2: The descendant concepts of the WordNet concept are compared to those of the ontology concept. If at least two of them share the same label, the concept in WordNet is chosen. Otherwise, proposition 3 is applied.
- Proposition 3: The ancestor concepts of a candidate concept are analyzed according to proposition 1. If the proposition is verified, the candidate concept is chosen for the mapping, if it is not the case, the concept is not disambiguated.

Concerning the identification of the generic concepts, the most specific ancestors of the mapped concepts are proposed to represent the generic concepts of the ontology. Examples of generic concepts extracted for the astronomy domain are:

"instrumentation": an artifact (or system of artifacts) that is instrumental in accomplishing some end.
"phenomenon": any state or process known through the senses rather than by intuition or reasoning
"natural object" : an object occurring naturally; not made by man

The generic concepts are proposed to the domain experts who judge their relevance. They are integrated in the ontology by the following process.

Integrating Generic Concepts In The Ontology
In order to link the level 0 concepts of the current ontology to the generic concepts previously extracted, an « is a sub-class » relation is defined between the level 0 concepts and the generic concepts extracted by the mapping.

Note that if for a given level 0 concept no generic concept has been extracted (no possible mapping in the generic ontology), the mapping to the generic concepts has to be done manually.

Structuring These Concepts: Associative Relations Extraction

The second stage in formalizing the ontology structure is to define the associative relations between concepts. This implies identifying a semantic relation between two concepts and labeling this relation. The process we propose relies on labeling the possible semantic relations that can be defined between pairs of generic concepts and using those labels to disambiguate the associative relations between concepts that can be extracted from the "related term" thesaurus relations.

Specifying Relations Between Generic Concepts

The specification of semantic relations between generic concepts is based on the proposition of relations associated to each pair of concepts by an automatic syntactic analysis of the reference corpus. These propositions form the basis of the manual definition of relations.

The context of the labels of each concept is extracted by the syntactic analysis of the reference corpus. By « context », we mean the phrase depending syntactically on each label (objects and subjects of verbs in which the labels appear). These contexts are then grouped according to the generic concepts they are linked with. The terms which occur frequently in the grouped contexts are retained to characterize the generic concept and are proposed as labels for the associative relations that the concepts linked to the generic concept may have. An example to illustrate this idea is *instrumentation* in the astronomy ontology. The terms occurring most frequently are the verbs « observe » and « measure » indicating that the astronomical instruments are used to observe and measure the other concepts of the domain.

Semantic relations are defined between each pair of generic concepts. A 2D matrix is built. This matrix contains, in lines and columns, all the different types of generic concepts identified manually based on the preceding propositions. Each cell contains the possible relations. An expert in the domain can thus identify the relations which link the pairs of generic concepts and add the labels he has chosen in the cell of the matrix. Figure 5 presents a sample of the matrix defined for our case study.

	Property	**Phenomenon**	**Event**	**Science**
Property	*Influences / Is influenced by* *Determined by / Determines* *Exclude* *Has part / Is part*	*Is a property of* *induces*	*Is a property of* *induces*	*Is studied by*
Instru-mentation	Makes Observes	*Observes* *Measures*	*Observes* *Measures*	*Is Used to studied*

Figure 5: Sample of the matrix defining the possible labels of semantic relations between generic concepts

This mechanism is a first step in helping experts determine relation labels. Experts have to validate or reject the propositions made according to verbs extracted in the context. The large number of verbs leads to numerous propositions giving experts many labels to verify. This step of our methodology can be considered as the most time-consuming as far as the experts are concerned. More sophisticated analyses are currently being considered.

Disambiguating Thesaurus "Related Term" Relations

The vague thesaurus « *related term* » relations are first transcribed in the ontology. Two terms linked in the thesaurus will thus lead to an association between the concepts for which they are labels in the ontology. This association is then labeled, thanks to the relations identified in the matrix between the generic concepts associated with the concepts. For example, the relation identified between the generic concepts « instrumentation » and « natural object» being the relation « *observes* », the thesaurus « related term » relation between « *coronagraph* » and « *solar corona* » (concepts derived from these two generic concepts) is modified into the relation « *coronagraph* » « *observes* » « *solar corona* ». If several semantic relations are identified, the choice is left to the domain expert.

Rule (R5) specifies the process.

Let ta_1 and ta_2 be two generic concepts with $ta_1 \in C_{Onto}$ and $ta_2 \in C_{Onto}$
Let $r,r' \in R_{Onto}$ with σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ and $r(ta_1,ta_2)$ with $G^{-1}(r)$ specified in the domain
If $r'(c_1,c_2)$ with $c_1 \in C_{Onto}$ and $c_2 \in C_{Onto}$ and c_1 « is sub-class of » ta_1 and c_2 « is sub-class od » ta_2 and $G^{-1}(r')$=« is linked to »
then $G^{-1}(r') \in G^{-1}(r)$

(R5)

MINING A TEXTUAL CORPUS TO ENRICH THE ONTOLOGY

Documents from the reference corpus are used to update the knowledge represented in the ontology. This update leads first to the definition of new semantic relations between existing concepts (first section). We also propose to extract new terms that were not previously in the ontology lexicon (second section) and to situate them according to the existing knowledge (section 5.3).

Detecting New Associative Relations Between Ontology Concepts

Contrary to approaches in the literature which aim solely at transforming a thesaurus into an ontology through the knowledge represented in it, we aim at establishing new associative relations between concepts by analyzing the textual documents of the domain (cf rule R6).

Using the previously established matrix, new relations can be found between the ontology concepts. For this, the syntactic context of the different concept labels is analyzed. The context is defined from the terms frequently co-occurring near the ontology concept labels.

When a concept label appears in the context of the label of another concept and the two concepts are not linked by a relation in the ontology, a relation is proposed. The label of this relation is established through the matrix defined above by considering the two generic concepts linked to the two concepts.

For example, in the context of the label « *luminosity* » referring to the concept of the same name, the label *« galaxy »* corresponding to the concept *« galaxy »* is found. As these concepts are of the types *« property »* and *« natural object »*, the relation « has a » is proposed between *« galaxy »* and *« luminosity »* (cf figure 5). As no relation has been previously established between the two concepts, the new relation is added to the ontology.

Let ta_1 and ta_2 be two abstract types with $ta_1 \in C_{Onto}$ and $ta_2 \in C_{Onto}$
Let $r,r' \in R_{Onto}$ with σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ and $r(ta_1,ta_2)$ with $G^{-1}(r)$ specified in the domain
If $r'(c_1,c_2)$ is extracted by the corpus analysis with $c_1 \in C_{Onto}$ and $c_2 \in C_{Onto}$
then $G^{-1}(r') \in G^{-1}(r)$

(R6)

Adding New Terms In The Ontology

In order to extract new domain terms, we consider two different weighting measures. The measures are complementary and aim at selecting from the set of terms extracted by the syntactic analyzer those that are most relevant.

The first weighting measure is the overall occurrence frequency of a term in the corpus. The frequently used terms and thus the most general ones are extracted. The terms considered are noun phrases, stop words such as articles or common verbs will thus not be extracted. The formula used is the following:

$$globality(term,corpus) = tf_{term,corpus} \quad \textbf{(1)}$$

where $tf_{term,corpus}$ represents the occurrence frequency of the term in the corpus

The second weighting measure aims at extracting specific terms of the corpus. It relies on tf.idf that is used in Information Retrieval to detect discriminating terms. It promotes terms appearing frequently in a document but rarely in the rest of the corpus. In order to apply this measure to term extraction, the measure is considered by the average td.idf obtained by each term on all the documents.

$$specificity(term,corpus) = \underset{\{documenti\} \in corpus}{average} (tf_{term,documenti} \times idf_{term}) \quad \textbf{(2)}$$

$$idf_{term} = \log(\frac{N}{f_{term}}) + 1$$

where $tf_{term,document}$ represents the occurrence frequency of a term in a document and f_{term} corresponds to the number of documents containing this term

According to a threshold, the terms extracted by measures (1) and (2) are proposed to be added to the ontology lexicon (R7 and R8).

If $t \in L_{corpus}$ and globality(t)>threshold
then $t \in L_{COnto}$
(R7)

If $t \in L_{corpus}$ and specificity (t)>threshold
then $t \in L_{COnto}$
(R8)

Figure 6 and 7 provide example of terms extracted by the two measures.

column density
high resolution
globular cluster
white dwarf
binary system
soft X ray
power law

Figure 6 Frequent terms of the astronomy domain not stated in the ontology

Yarkovsky force
Relativistic gravity
Suprathermal electron
Halpha knot
Penumbral wave
Mean free path
Integral magnitude
Mixing layer
stellar population

Figure 7 Specific terms of the astronomy domain not stated in the ontology extracted from the thesaurus

Adding New Semantic Links Between Concepts

The new terms extracted by the previous measures have to be integrated in the ontology. The integration process we propose relies on the comparison of the words composing the extracted terms and the terms forming the concept labels in the ontology. More specifically, we consider for each extracted term the main word (i.e. the word on which the remaining words depend syntactically) and the remaining ones. For example, in a nominal phrase "column density" the main word is "density" as "column" is a noun complement.

Three different processes are applied.

If the main word and the rest of the term refer to two concept labels in the ontology, the corresponding concepts and their generic concepts are extracted. The new term is integrated in the ontology by the creation of an associative relation is then created between the two concepts. The relation is labeled on the basis of the relations defined between the two generic concepts. R9 expresses this process.

If only the main word corresponds to a concept label, the new term is proposed to be the label of a new concept that will be subsumed by the concept referenced by the main word. As the new term defines information on the main word, it specifies the meaning of the concept referenced by the main word. (cf rule R10).

Let $t \in L_{corpus}$ to be added to L_{COnto} , $ta_1 \in C_{Onto}$ and $ta_2 \in C_{Onto}$
If main_term(t) $\in L_{COnto}$ with F(main_term(t)) « sub class of » ta_1 and remaining(t) $\in L_{Conto}$ with
F^{-1}(remaining(t))« sub class of » ta_2
and $r \in R_{Onto}$ with σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ and $r(ta_1,ta_2)$ with $G^{-1}(r)$ specified in the domain
then $r' \in R_{Onto}$ avec $G^{-1}(r') \in G^{-1}(r)$
(R9)

Let $t \in L_{corpus}$ to be added to L_{COnto} ,
If main_term (t) $\in L_{COnto}$ and remaining(t) $\in L_{Conto}$
then $t \in L_{COnto}$ with F(t)=c and c « sub-class of » F(tmain_term(t))
(R10)

CASE STUDY: BUILDING AND USING AN ONTOLOGY IN THE ASTRONOMICAL FIELD

We have applied the proposed method to the transformation of a thesaurus into a lightweight ontology in the field of astronomy. The IAU thesaurus (http://www.mso.anu.edu.au/library/thesaurus/) was built to standardize the terminology of astronomy. Its use was to help librarians index and retrieve catalogues and scientific publications of the domain. Its building, requested by International Union of Astronomy in 1984, was finished in 1995. Its transformation into an ontology has been done in the framework of the French project Data Mass in Astronomy (http://cdsweb.u-strasbg.fr/MDA/mda.html). We present in section 6.1, the results obtained by applying the different rules presented in the article. In section 6.2, we give an example of how the constructed ontology has been integrated to a science-monitoring task.

Evaluation Of The Proposed Method

Two corpuses of the domain have been considered. The documents they contain are abstracts of articles published in the international journal Astronomy and Astrophysics (www.edpsciences.org/aa). The first corpus is composed of articles published in 1995. The use of this corpus aims at capturing the implicit knowledge of the domain not stated in the thesaurus when it was built. The second corpus contains articles published in 2002. This corpus has been chosen to enable the update of the domain knowledge from recent documents. Domain experts have validated that both corpuses describe the knowledge to be represented in the ontology.

The protocol defined to evaluate the transformation relies on presenting the results obtained by the different rules to two astronomers who accept or reject the propositions.

We present here the most significant results.

- Extraction of concepts

Terms in the thesaurus	Concepts created	Concept validated
2957	2547	85%

- Definition of the new abstract level in the ontology

Mapping of top level concepts to concepts of the generic ontology (WordNet)	
Top concepts that have candidate concepts in the generic ontology	72%
Top concepts for which the disambiguation process was successful	65%
Extraction of generic concepts	
Number of generic concepts extracted	19
Number of generic concepts validated	74%
Integration of generic concepts in the ontology	
Top level ontology concept correctly associated to a generic concept	89¨%

- Extraction of associative relations

Disambiguation of the associative relations extracted from the thesaurus "is related to" relation	
Number of relations evaluated	49
Correctly labeled relations	84%

- Ontology update

Extraction of new associative relations between concepts	
Number of proposed relations	74
Validated relations	92%
Validated labels	87%
New terms extraction	
Percentage of correct global terms extracted	72%
Percentage of correct specific terms extracted	62%
New terms integration to the ontology	
Percentage of correct new concepts extracted	100%
Percentage of correct new relations extracted	62%
Validated labels	87%

The evaluation of our method in the field of astronomy has demonstrated its interest. The aim of our method is not to be automatic but to facilitate the creation of ontologies. Note that for each step the task of experts is to confirm or reject propositions and not to give results. For this, our method is efficient. The identification of domain concepts and their labels is highly relevant (85% of the concepts are validated). It leads to the definition of generic concepts that are mostly validated and facilitates their integration in the ontology. Associative relations extracted from the thesaurus are validated for 84%. Thanks to the mining of the reference corpus, relevant semantic relations between existing concepts are added and new terms integrated by means of the creation of new concepts or new relations.

Using Ontologies For A Science Monitoring Task

Ontologies are used for semantically indexing documents. This relies on the intuition that the meaning of textual information (and the words that compose the document) depends on the conceptual relations between the objects to which they refer rather than on the linguistic and statistical relations of their content (Haav and Lubi, 2001).

Semantic indexing is done in two steps: identifying document concepts within documents and weighting those concepts. The weighting process uses the relations between concepts identified in the documents according to the structure of the ontology. For more precise details, see (Hernandez et al., 2005).

Based on this semantic indexing, we propose to take the ontology into account to provide the user access to the information contained in a corpus. The interface is developed to support OWL ontology visualization. A snapshot is presented in figure 8. With the interface, it is possible to visualize both the domain ontology linked to the content of documents and the specific meta-data the user is looking for.

The user explores the corpus according to the specific theme of the domain (in our case astronomy).

This possibility is illustrated in figure 9. On the right hand side of the screen, the domain ontology is shown with its concepts and relations. The concepts present in the corpus are highlighted; they can be interpreted in their context as their relations to other concepts are represented. By clicking on a concept, the user can find the articles treating the theme and the researchers working on this theme. When a researcher working on the theme is selected, the windows are automatically updated and the information on this researcher is presented. In the same way, when the user is interested in accessing the article dealing with this subject, he selects it. The windows are updated: a pop-up window containing the article appears and the information linked to this article is represented as in figure 8. The information known about this article is presented on the right hand side of the screen and all the themes of the domain treated in the article are presented in the astronomy ontology frame. The astronomy ontology

concepts found in the document are presented in red on the left-hand side of the window (the article of reference 124 deals with comet and solar system). This makes it possible to evaluate rapidly if the document treats the themes in which the user is interested, who the authors of the article are, when it was published and where.

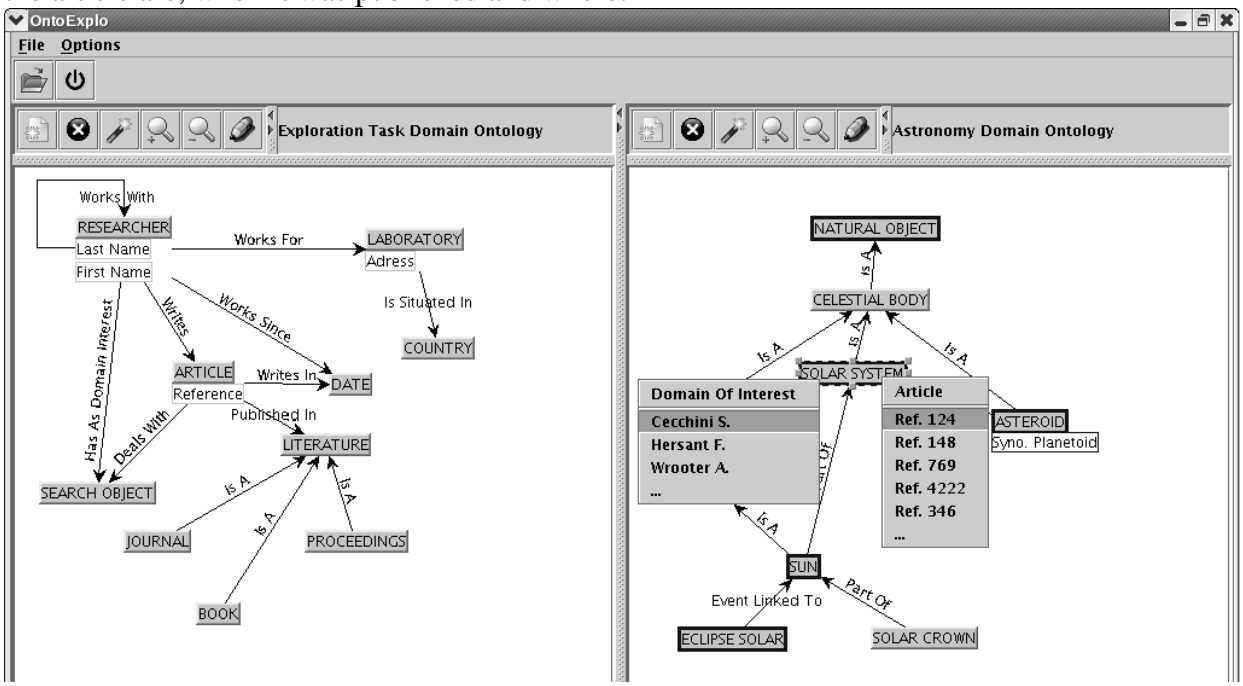

Figure 8. Visualization through the astronomy ontology

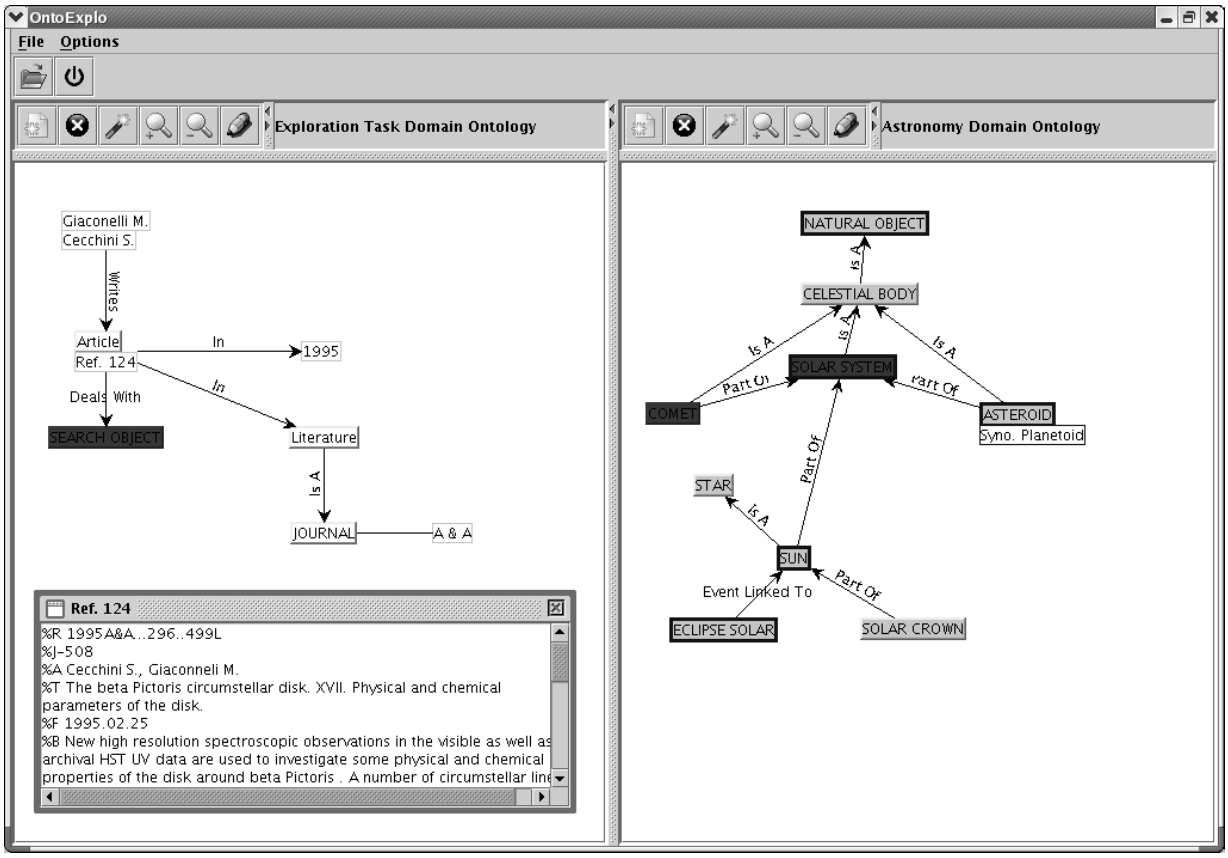

Figure 9. Visualization of the knowledge learnt for an article

CONCLUSIONS AND FUTURE WORKS IN THE DOMAIN

The representation of textual contents is an issue to be addressed if systems are to handle documents efficiently. The assumption according to which documents contain all the information necessary to their representation is no longer enough. However, a more semantic indexing implies the use of terminological resources or domain knowledge. In this chapter, we have proposed a methodology that produces this type of linguistic resource in a formal way, based on ontologies. This methodology applies when a thesaurus and a textual corpus on the domain are available. The process is based on the following main steps: extraction of initial knowledge from the thesaurus itself (basically terms, concepts and taxonomic links), enrichment of the resulting ontology based on text mining (disambiguation of existing relations between concepts, extraction of new terms and new relations from texts). We thus define an incremental process that can be applied to up-date existing ontologies as well. This process is automated as much as possible and the work of experts is limited, consisting only in validating the results (generic concepts, relations labels). Indeed, our methodology and its implementation is less demanding with regard to experts' time (Soergel et al., 2004) (Wielinga et al, 2001). An important result is that relations between concepts are semi-automatically extracted and a label is associated to them.

This work is a contribution to the semantic web which structures the web in a way that it is meaningful not only to computers but also to humans. (Lu et al., 2002) consider that one of the main challenges of developing the Semantic Web is through ontologies. To meet this challenge new methods and tools are needed in order to facilitate this process.

REFERENCES

Aussenac-Gilles N., Biébow B., Szulman S. (2000) Revisiting ontology design : a method based on corpus analysis, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), R Dieng, O. Corby (Eds.), 172-188.

Aussenac-Gilles N., Mothe J. (2004) Ontologies as Background Knowledge to Explore Document Collections, In Proceedings of RIAO, 129-142

Bozsak E., Ehrig M., Handschuh S., Hotho A., Maedche A., Motik B., Oberle D., Schmitz C., Staab S., Stojanovic L., Stojanovic N., Studer R., Stumme G., Sure Y., Tane J., Volz R., Zacharias V. (2002) KAON - Towards a large scale Semantic Web, In Proceedings of the 3rd International Conference on E-Commerce and Web Technologies, volume 2455, 304-313

Bourigault D., Fabre C (2000) Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaire, 25, Université Toulouse le Mirail, 131-151, 2000

Chaumier J. (1988) Le Traitement linguistique de l'information. Entreprise moderne d'éd., ISBN 2-7101-0684-1

Crouch CJ. and Yang B. (1992) Experiments in automatic statistical thesaurus construction, Conference on Research and Development in Information Retrieval (SIGIR), pages 77-88, 1992.

Ding Y., Foo S., Ontology Research and Development: Part 1 – A Review of Ontology Generation, Journal of Information Science 28(2)

Englmeier K., Mothe J. (2003) IRAIA: A portal technology with a semantic layer coordinating multimedia retrieval and cross-owner content building, In Proceedings of the International Conference on Cross Media Service Delivery, Cross-Media Service Delivery Series, The International Series in Engineering and Computer Science, V. 740, 181-192

Fensel S. B. (1998) Knowledge Engeeneering : Principles and Methods. Data and Knowledge Engineering, 25, 161-197

Fischer D. H. (1998) From Thesauri towards Ontologies?, In Structures and Relations in Knowledge Organization : Proceedings of the 5th International ISKO Conference, W.M. Hadi, J. Maniez, S. Pollitt (Eds.),Würzburg: Ergon, 18-30
Foskett D.J. (1980) Thesaurus, In Encyclopedia of Library and Information Science, A. Kent, H. Lancour (Eds), 416-463
Gal A., Modica G., Jamil H.M. (2004) OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources, In Proceedings of the 20th International Conference on Data Engineering, IEEE Computer Society
Gangemi A., Guarino N., Masolo C., Oltramari A., Schneider L. (2002) Sweetening Ontologies with DOLCE, Proceedings of the International Conference on Knowledge Engineering and Knowledge Management., 166-181
Grefenstette G. (1992) Use of syntactic context to produce term association lists for retrieval, Conference on Research and Development in Information Retrieval (SIGIR), 89-97
Haav H.M., Lubi T.L. (2001) A Survey of Concept-based Information Retrieval Tools on the Web, In Proceedings of the 5th East-European Conference ADBIS, Vol 2, 29-41.
Hahn U., Schulz S. (2004) Building a Very Large Ontology from Medical Thesauri, Handbook on Ontologies, S. Staab, R. Stuber (Eds.) 133-150
Harman D. (1992) The DARPA TIPSTER project, In SIGIR Forum, volume 26(2), 26-28
Hearst M.A. (1992) Automatic acquisition of hyponyms from large text corpora, In Proceedings of the 14th International Conference on Computational Linguistics
Hernandez N., Mothe J., Poulain S. (2005) Accessing and mining scientific domains using ontologies: the OntoExplo System, Poster, In Proceedings of The 28th Annual International ACM SIGIR, 607-608
Lassila O., McGuiness D. (2001) The role of frame-based representation on the semantic Web, Rapport technique KSL-01-02, Knowledge Systems Laboratory, Stanford University
Lu S. and Dong M. and Fotouhi F. (2002) The Semantic Web: opportunities and challenges for next-generation Web applications, http://informationr.net/ir/7-4/paper134.html
Guarino N., Carrara M., Giaretta P. (1994) Formalizing ontological commitments, In Proceedings of the AAAI conference
Gómez-Pérez A., Fernandez M., de Vicente A.J. (1996) Towards a Method to Conceptualize Domain Ontologies, In Proceedings of the European Conference on Artificial Intelligence (ECAI'96), 41-52
Maedche A., Staab S. (2001) Ontology Learning for the Semantic Web, IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2)
McGuinness D.L (2004) van Harmelen F., OWL Web Ontology Language Overview, W3C Recommendation http://www.w3.org/TR/owl-features/, 10 February 2004.
Miles A., Brichley D. (2005) SKOS Core GuideW3C Working Draft 10 May 2005, http://www.w3.org/TR/swbp-skos-core-guide/
Mills D. (2006) Semantic waves 2006, http://www.ift.ulaval.ca/~kone/Cours/WS/WS-PlanCours2006.pdf
Miller G.A. (1988) Nouns in WordNet, In WordNet, An Electronic Lexical Database C. Fellbaum (Ed), 23-46, MIT Press
M. Porter (1980) An algorithm for suffix, Stripping Program, 14(3), 130-137
Qiu Y. et Frei H.P. (1993) Concept based Query Expension, Conference on Research and Development in Information Retrieval (SIGIR), 160-169
Robertson S. E., Sparck Jones K. (1976) Relevance weighting of search terms, Journal of the American Society for Information Sciences, 27 (3), 129-146
Salton G. (1971) The Smart Retrieval System, Prentice Hall, Englewood Cliffs, NJ
Smith M.K., Welty C., McGuinness D.L., OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-guide/

Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. (2004) Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information, Volume 4 Issue 4
Tudhope D., Alani H., Jones C. (2001) Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, 1-8(41)
Uschold M., King M. (1995) Towards a Methodology for Building Ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995)
Wielinga B., Schreiber G., Wielemaker J., Sandber J.A.C. (2001) From thesaurus to ontology, In Proceedings of the International Conference on Knowledge Capture

FUTURE RESEARCH DIRECTIONS

Mining for information on the semantic web refers to various points of views (Corby et al. 2004): that of designing ontologies which aim at representing a field of knowledge, that of semantically indexing resources according to these ontologies, finally, that of providing access to resources for a user or an engine. The work we present focuses on the first point of view.

Considering the same point of view, one of the challenges is the evolution of domains and related needs. Domain knowledge evolves with time. Now that methods have been defined to help ontology building, research should take into consideration how ontology can evolve. It implies populating existing ontologies with new instances that are extracted from recent domain resources, restructuring ontologies when new concepts are identified or when knowledge become obsolete. Ontology evolution is also linked to the knowledge needed by information systems to better interact with users. When applications are modified or when user's expertise of the application or the field changes, the ontology has to be restructured according to what is really useful. The PASCAL Ontology Learning Challenge aims at incouraging and evaluating work in such directions.

Semantic Web research incorporates a number of technologies and methodologies defined in different fields including Artificial Intelligence, Ontology Engineering, Human Language Technology, Machine Learning, Web Services, Databases, Distributed Systems, Software Engineering, Human Computer Interaction and Information Systems. One of the challenges is to make these communities work together in order to integrate the advances in the different fields. Moreover, many application domains have been investigated (e-Business, e-Health, e-Government, Multimedia, e-Learning, Bio-Pharmacy) but scalable applications is still a challenge when considering the 11.5 billion pages of indexable Web.

Another challenge is that ontologies are yet build in order to state the consensual knowledge of the domain. In order to be applied on large scale document repositories where many groups of users interact, processes should be developed to add descriptions to ontologies in order to define in which context and for whom the ontology is useful. This last point is part of the aims of the pragmatic web which is envisioned to be the new evolution of the semantic web.

ADDITIONAL READING

Brewster C., Alani H., Dasmahapatra S., Wilks Y. (2004) Data driven ontology evaluation, In Proceedings of 4th International Conference on Language Resources and Evaluation
Chebotko A., Lu S., Fotouhi F. (2006) Challenges for Information Systems Towards The Semantic Web, In the SIGSEMIS journal, http://www.sigsemis.org/?p=9
Chrisment C., Dousset B., Dkaki T. (2007) Karouach S., Mothe J., Combining Mining and Visualization Tools to Discover the Geographic Structure of a Domain, In Computers Environment and Urban Systems Journal, to be published in 2007

Corby O., Dieng-Kuntz R., Faron-Zucker C. (2004) Querying the Semantic Web with Corese Search Engine, In proceedings of the European conference on artificial intelligence, 705-709
Desmontils E., Jaquin C.(2002) Indexing a Web site with a terminology oriented ontology, The Emerging Semantic Web, I. Cruz S. Decker, J. Euzenat, D.L. McGuinness (Eds.), IOS Press, ISBN 1-58603-255-0, 181-197
Englmeier K., Mothe J. (2003) IRAIA: A portal technology with a semantic layer coordinating multimedia retrieval and cross-owner content building, In Proceedings of the International Conference on Cross Media Service Delivery, Cross-Media Service Delivery Series, In proceedings of the International Series in Engineering and Computer Science, V. 740, 181-192
Guha R.V., McCool R., Miller E. (2003) Semantic search, In Proceedings of the 12th International World Wide Web Conference, 700-709
Hernandez N., Mothe J. (2004) An approach to evaluate existing ontologies for indexing a document corpus, In Proceedings of The Eleventh International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA) -Semantic Web Challenges-, 11-21
Hernandez N., Mothe J., Chrisment C., Egret D. (2007) Modeling context through domain ontologies, Journal of Information Retrieval, Special issue Contextual Information Retrieval, to be published in 2007
Jarvelin K., Ingwersen P. (2004) Information seeking research needs extensions towards tasks and technology, Information Retrieval, 10 (1), 212
Kiryakov A., Popov B., Terziev I., Manov D., Ognyanoff D. (2004) Semantic annotation, indexing, and retrieval, Journal of Web Semantics, 2(1)
Lozano-Tello A., Gómez-Pérez A. (2004) ONTOMETRIC: A Method to Choose the Appropriate Ontology, Journal of Database Management, 15(2)
Mothe J., Chrisment C., Dousset B., Alaux J. (2003) DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets, Journal of the American Society for Information Science and Technology, Vol 54 (2), 650-659
Shaban-Nejad A., Baker C. J. O., Haarslev V., and Butler G., (2005), The FungalWeb Ontology: Semantic Web Challenges in Bioinformatics and GenomicsSemantic Web Challenges
Studer R., Benjamins R., Fensel D. (1998) Knowledge Engineering: Principles and Methods, Data and Knowledge Engineering, 25(1-2) 161-197
Uschold M. (2003)Where are the semantics in the semantic Web?, In AI Magazine, 24(3), 25-36, ISSN:0738-4602
Vallet D., Fernández M., Castells P. (2005) An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, 455-470