Modeling context through domain ontologies

Nathalie Hernandez	Josiane Mothe	Claude Chrisment	Daniel Egret
IRIT-UPS	IRIT-UPS, IUFM	IRIT-UPS	Président de
118 route de Narbonne	118 route de Narbonne	118 route de Narbonne	l'Observatoire de Paris
31062 Toulouse	31062 Toulouse	31062 Toulouse	61 avenue de
Cedex 9, France	Cedex 9, France	Cedex 9, France	l'Observatoire
hernande@irit.fr	mothe@irit.fr	chrisment@irit.f	75014 Paris
			egret@obspm.fr

Abstract

Traditional information retrieval systems aim at satisfying most users for most of their searches, leaving aside the context in which the search takes place. We propose to model two main aspects of context: the themes of the user's information need and the specific data the user is looking for to achieve the task that has motivated his search. Both aspects are modeled by means of ontologies. Documents are semantically indexed according to the context representation and the user accesses information by browsing the ontologies. The model has been applied to a case study that has shown the added value of such a semantic representation of context.

Keywords: ontology, task, document representation, semantic indexing, browsing interface

1 Introduction

Information Retrieval (IR) can be defined as an activity which aims at localizing and delivering granules of information (document, part of document, set of documents) to a user, according to his information needs. Many pragmatic suppositions have been made in the literature to facilitate the engineering of systems that can deal with huge amounts of information (Jones 2004). One of these consists in proposing systems that satisfy most users for most of their searches (Spark Jones 1999). The kernel of these systems is independent of the context. The mechanisms focus on the representations of documents and queries and on their mapping and leave aside the modeling of the context linked to the user's search. In order to propose systems providing more precise answers to the

user's need, the domain of contextual IR has recently become a priority (Allan 2003). The aim is to put the user back into the centre of IR models by expliciting some elements of the context that can affect the system's performances. Context refers to the knowledge linked to the user's intention (task, perception of the task, type of information needed), to the user himself (a priori knowledge, profile), to his environment (material environment, culture), to the domain of his information need (corpus nature, domain treated), and to the characteristics of the system (document representation, query/document mapping, strategies for accessing documents, visual interface). Taking context into account implies both identifying and modeling the different aspects which are useful to specify the user's information need and then integrating them into IR methods and processes.

Taylor (Taylor 1968) dissociates two distinct but interwoven parameters for defining information needs. The first is the theme or the subject of the need which determines what the information will treat. The second concerns the task or the user's situation and this conditions the reasons for which the information is searched and how it will be used. The majority of IR Systems (IRS) focuses on the first parameter, the theme (Freund 2005). Moreover, this aspect is taken only partially into account as systems settle for searching the granules containing the terms given by the user to specify the theme of his need, even though terms are often ambiguous and can refer to several themes.

Our work aims at better taking into account the theme of the need and the user's task by setting them in their knowledge domain. Domain knowledge can be represented through thesauri or ontologies. A thesaurus is based on the main terms of one or more fields organized by conventional relations, such as “is a broader term”, “is a related term” (AFNOR 1987). These relations lack semantics and expressiveness. The “is a broader term” relation can imply a degree of genericity/specificity but can also refer to “is an instance of” or “is a part of” relations. The associative relation “is a related term” is vague and implies different semantic relations (“is a characteristic”, “produces”, “belongs to”…) (Tudhope 2001). This can be explained by the fact that thesauri have been built to help librarians index and retrieve documents by means of thematic categories, not to provide formal representations. Attempts have been made to integrate them in automatic processes (Baeza 1999) (Tudhope 2001), but as the implicit knowledge of librarians is not included, the thesaurus failed to improve system performances. An ontology corresponds to « an explicit and formal specification of a shared conceptualisation » (Studer 1998) and has a higher degree of formalization than a thesaurus. At the heart of the semantic web, ontologies are used to add a semantic layer to the actual web. They are references for communication between machines but also between machines and humans by defining consensually the meaning of objects firstly through symbols (words or phrases) that designate and characterize them, and then through a structured or formal representation of their role in the domain (Aussenac 2004)., Ontologies for IR can therefore play the role thesauri used to play by defining an indexing language relying on a more formalized knowledge that can be exploited in the retrieval process.

The framework of our study is that of documentary databases of a specific domain. Such databases are more and more common in scientific fields (e.g. medicine, astronomy) as well as in enterprises. We have chosen to base our method on domain rather than generic ontologies (WordNet (Miller 1988), DOLCE (Gangemi 2002)) as they are much more appropriate in this context. Specific domains are better covered as the representation is limited to terms, concepts and relations in the field. When restricting the interpretation of concepts to the specific domain context, the ambiguity of the terms composing the ontology is limited and finding concepts in the documents easier. Additionally, within the context of databases of a specific domain, users generally share a common language; using domain knowledge helps communication between users in the language of their community (Englmeier 2004).

Following (Freund 2005) we propose a model based on two ontologies which represent the two aspects of the context (search task and theme). The first ontology (theme domain ontology) specifies and structures the objects of the themes treated in the documents and their relations. The second (task domain ontology) specifies the data the user is interested in, according to the type of task he is engaged in; it is represented through meta-data.

The model is integrated in an IRS in two phases of the IR process. It is first integrated in the granule representation through a semantic indexing mechanism. This relies on the intuition that the meaning of textual information (and the words that compose the granules) depends on the conceptual relations between the objects to which they refer, rather than on the linguistic and statistic relations of their content (Haav 2001). Semantic indexing consists in searching for the concepts referenced in the granules (instead of terms) and weighting those concepts according to the semantic representativity of the granules. Secondly, it is integrated in a browsing process the user manipulates in order to access the document collection.

The proposed model and its integration in an IRS were applied to a case study in the field of astronomy in which Astronomers took part. In this domain, much effort is being put into elaborating "virtual observatories" that provide an optimal scientific use of all the data available in this field. These Observatories notably need to improve researchers' access to textual digital documents produced by the scientific community. The astronomers are particularly interested in performing science monitoring tasks to analyze the research themes of the different actors of the domain (researchers, laboratories, journals …). To access scientific publications, they generally use the existing ADS[1] (Astrophysics Data System) server which allows them to specify their need only by term queries. Our proposition has thus been applied to this task in this particular domain in order to evaluate the usefulness of specifying the search context with domain knowledge. In the following sections, the case study is used to illustrate our approach.

The article is organized as follows. First, we present the semantic model of the search context by defining and illustrating the task domain ontology and the theme domain ontology. Then, we

[1] http://cdsads.u-strasbg.fr/

describe how the model is integrated in the IRS when indexing the documents and accessing information. We illustrate our approach by the description of the interface that has been used for its validation. Finally, we present an informal evaluation of our approach.

2 Related works

The variability of context aspects in IR makes them difficult to model. Ontologies are, however, one way of modeling context by specifying the domain knowledge linked to the aspects considered. Semantic indexing can be used to integrate the model when representing and accessing granules.

In IR, semantic indexing is done in two steps: identifying document concepts within granules and weighting those concepts. The weighting process that, according to the definition of semantic indexing, should use the relations between concepts identified in the granules does not usually consider the structure of the ontology, or only partially. This lack of semantic consideration leads authors to combine traditional indexing with semantic indexing. In (Baziz 2005) (Vallet 2005) (Mihalcea 2000), concepts are only weighted according to an adaptation of the traditional measure of td.idf (Robertson 1976). In (Desmontils 2002), the representativity of a concept is calculated according to whether, in the structure of the ontology, it is linked to other concepts found in the granule. However, the method proposed only takes into consideration taxonomic relations. For example, if the concepts “information retrieval” and “information retrieval systems” are found in a granule, the relation between those two concepts will be considered although their semantic relation is trivial. Other relationships such as meronomical (“is a part of”) or temporal relations (e.g. “appears before”) are not taken into account.

Semantic indexing is also a process used in the semantic web. Here, it is not only a case of indexing document content thanks to ontological resources but also of representing the resources present in the semantic web by generating the corresponding meta-data. The aim is to add to the web a formal structure and semantics (through meta-data and ontological knowledge) in order to improve the management of and access to information. The approach relies on ontologies modeling objects of the world through the actors and the entities that the documents constitute and contain (Guha 2003). In the context of the semantic web, concepts and instances of concepts are mainly extracted from granules but not weighted (Kiryakov 2004). It takes advantage of the formal representation of the ontology to infer concepts that can be referenced in the user’s query, but not explicitly stated in granule contents. However, when several granules match the query, they cannot be ranked according to their similarity to the query.

We promote an approach in which these two principles are combined: semantic indexing including meta-data descriptions and ranking results. Some other systems also integrate the two approaches by using a single ontology representing meta-data and content or theme concepts. In (Benjamins 1999), yellow pages about researchers are annotated manually according to an ontology describing the researcher, laboratory, publications and research themes. In (Vallet 2005), the ontology

is defined according to abstract classes such as meta-data, documents and theme. In (Stuckenschmidt 2004), biological articles are indexed using a domain thesaurus (containing hierarchical relations between terms and semantically vague associative relations, such as "is a drug linked to", "is a disease linked to"). The automatically extracted index and the meta-data, such as the title of the article, name(s) of authors, are represented with an ontology. In these approaches, no distinction is made between the knowledge used for representing the theme of the need and the task. From our point of view, this kind of system lacks reusability. Even if the content of the documents and the associated meta-data are treated, the fact that the two aspects are not dissociated implies that these systems would not easily be applicable to other corpuses. Specific corpus meta-data will not be taken into account if they are not contained in the domain ontology. Moreover, the system will not be able to deal with a user wanting to specify his need according to what he is searching for in the documents dealing with particular themes, i.e. the documents relevant for his task. He can either specify his need according to domain terms or to meta-data, but cannot combine both aspects.

Using ontologies to access information but not separating both aspects has several drawbacks. Cat-a-cone (Hearst 1997) for example is a system in which documents are retrieved by browsing the Mesh hierarchy of concepts. Even if hierarchical organization is an intuitive way of structuring information, the interface only enables the visualization of the corpus according to theme concepts describing the content of documents. Leaving aside the meta-data associated to the documents implies a waste of information, especially nowadays with the growing interest of the semantic web. Moreover, hierarchies of concepts rely only on taxonomic relations. The browsing context is semantically richer and less ambiguous when different semantic relations are considered. This has been pointed out in the work carried out in (Croft 1987) in which the browsing interface displays semantic relations. Domain knowledge is used to construct stereotype users' models. User models are composed of terms, concepts and semantic relations that are designated by experts to be relevant for a specific user's need. The relations are "synonymy", "generalization", "is an instance of" "is a part of" and "cross references" (note that the last relation refers to the "is related term" relation that can be found in thesauri and that is semantically vague). Various heuristics are used by the browsing interface to decide which links to show the user to help him formulate his query according to the stereotype model he is related to. In this system, domain knowledge is only used to help formulate queries. It is not taken into account when indexing and clustering documents. The system could be improved by using an ontology thereby taking into account meta-data and could rely on clustering techniques based on semantics such as those used in the works cited below.

In (Stuckenschmidt 2004), corpus exploration and access to documents are done by clustering the documents that share the same ontology concepts and instances. Through an interface, the user can access the clusters and concepts which have led to the clustering by entering a term referencing a concept. He can also display the terms occurring most frequently in the cluster and decide to focus on one of them to continue his browsing. A similar approach is proposed in (Seeling 2003). The system relies on an ontology for computing similarity between documents which share the same meta-data of

a specific domain. In the two latter approaches, the ontologies on which they are based are not presented to the user in order to explicit the context of the search. Providing the user with the knowledge representation linked to the search context gives him a global view of the available information in the corpus and helps him interpret the context according to his knowledge on the domain.

The user specifies his need according to the information he thinks he will find in the corpus (Turtel 1991). One way for him to be able to specify his need in function of the themes actually treated is to present them according in the general context of the domain(s) treated. The context must then reflect the themes linked to the domain knowledge and situate the themes in relation to each other. A user does a search task to access precise data satisfying his need. The context must explicit the information sought and the reasons why and how it can be interpreted (Jones 2004). In our model, this is done through the meta-data associated to the documents that interest the user. The meta-data and their links, explicit or implicit, are specified according to the objects of the search and their role in the task. For example, in astronomy, researchers may be interested in exploring scientific journals for two different tasks. When exploring a corpus for science monitoring on the activities of a research laboratory, they are interested in finding, from the granules content, the names of the researchers, the type of publications, their themes of research and the different correlations between these actors. When performing a bibliographical task, they are looking for specific observations and measurements on domain objects and their evolution. The corpuses explored in both tasks are the same, but the specific data the astronomers are looking for differ. The themes remain the same but their integration in the search depends on the context of the task. In the same way, a task can be done on a corpus of another domain, the actors of the task (researchers, publications, if we take the previous example) are the same but their correlations to the themes and domain treated in the corpus are modified. To enable reuse on different tasks and corpuses, our model is based on two different ontologies: a task domain ontology linked to the user's purposes, and a theme domain ontology linked to the domains treated in the corpus. They are described in the following sections.

3 Semantic context model

3.1 Formalization

The context model relies on lightweight domain ontologies (Lassila 2001) and domain models. Both are presented and formalized in the following sections.

3.1.1 Lightweight ontologies

The use of ontologies is the extension of the traditional use of lexical resources in IR (Lassila 2001). Descriptors are no longer chosen from controlled vocabularies (or thesauri) but within ontologies that consensually represent terms, concepts and semantic relations of the domain. Document granules are indexed by concepts that represent their meaning rather than by words that may be ambiguous (Aussenac 2004). The approach uses lightweight ontologies composed of two

semiotic levels (Maedche 2002). The lexical level (L) covers all the terms or labels defined to designate the concepts and enables their detection within granule contents. The conceptual level defined in the structure (S) of the ontology represents the concepts and the semantics defined from the conceptual relations that link them. Characteristics of the relations, such as symmetry and transitivity, can also be considered.

The **structure** of an ontology is a tuple S: = {C, R, A, T, CAR_R, $\leq_C$, σ_R, σ_A, σ_{CAR}} where:

- C, R, A, T, CAR_R are disjoint sets containing concepts, associative relations, attribute relations, data types, and the characteristics of associative relations (synonymy, transitivity …)
- $\leq^C$: C x C is a partial order on C, it defines the hierarchy of concepts
  $\leq^C (c_1, c_2)$ meaning that c_2 "is a" c_1 is called a taxonomic relation
- σ_R : R $\rightarrow$ C x C is the signature of an associative (or non-taxonomic) relation
- σ_A : A $\rightarrow$ C x T is the signature of an attribute relation
- σ_{CARR} : R$\rightarrow$ CAR_R specifies the characteristic of an associative relation

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

- L^c and 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-1(r)={l / l$\in L^R$}

Note that F and F^{-1} are not mathematical applications. A concept can be defined by several terms and a term, when it is ambiguous, can reference several concepts.

The proposed formalism is implemented by using OWL specifications (McGuinness 2004). In this language recommended by the W3Cc, concepts are *owl:concepts* defined with label properties, associative relations are *owl:Objects* properties and attribute relations are *owl:Dataproperties*.

3.1.2 Domain model

The model of a domain is represented by an ontology as previously defined and by the instances that are associated to its concepts.

It is formalized by the tuple (O,I, V, f_C, f_T, f_R, f_A) with :

- O the lightweight domain ontology
- I the set of instances
- V the set of values of the data types
- f_C : I $\rightarrow$ C, a concept instantiation function
- f_T: V $\rightarrow$ T, a data type instantiation function

- $f_R : I \times I \rightarrow R$, an associative relation instantiation function
- $f_A : I \times V \rightarrow A$, an attribute relation instantiation function

3.1.3 Building ontologies

Many newly developed techniques require and enable the specification of ontologies (Staab 2004). Some methodologies (such as Methontology (Fernandez 1997), OntoClean (Guarino 2002)) rely on manual design guided by ontological considerations and collaborative work. Others (such as Terminae (Biblow 1999), Text-to-Onto (Cimiano 2005)) aim at facilitating construction by extracting knowledge from reference textual corpuses with natural language processing techniques. The latter are particularly adapted to IR as they enable the building of representations from the documents that are considered by the IRS. Some methods (such as TtoO (Mothe 2006) or (Soergel 2004)) are even better adapted as they rely both on the reengineering of thesauri into conceptual resources and on the updating of knowledge thanks to textual domain documents.

Building ontologies is however a time consuming task. Thanks to the growing interest in ontologies in the field of information systems, freely accessible ontologies are emerging. The methodology defined in (Hernandez 2004) makes it possible to evaluate whether a given existing ontology is suitable for indexing a document corpus.

In our approach, we consider that any methodology can be used to build the task and theme ontologies. For our case study, the task ontology was built manually using Methontology and took the astronomers and ontology engineer less than 30 hours. The theme ontology was built, in less than 10 hours, in another part of the project “data mass in astronomy”, and reused according to the analysis made by the methodology defined in (Hernandez 2004)..

3.2 Task domain context

3.2.1 Definitions

3.2.1.1 Ontology

A task domain ontology is composed of a set of concepts interesting the user for his IR task and a set of relations specifying their role in the task. It is chosen for a given task for a user or group of users in function of the information targeted. It can be either reused or elaborated for the task in question; the main point is that it is validated by the user and that its lexicon and its structure specify the conceptualization linked to the intended task. The task domain ontology is reusable for the same task on different corpuses.

This ontology is composed of

- a structure $S_{task} := \{ C_{task}, R_{task}, A_{task}, T_{task}, CAR_{Rtask}, \leq^{C}_{task}, \sigma_{Rtask}, \sigma_{Atask}, \sigma_{CARR} \}$
- a lexicon $L_{task} := \{L^{C}_{task}, L^{R}_{task}, F_{task}, G_{task}, \}$

For practical reasons, we will use the term « task ontology » to designate the « task domain ontology ».

3.2.1.2 Task domain model

The task model is defined by the task ontology described previously and by instances of its concepts that can be extracted from the documents and by the functions that instantiate these elements. The model is thus defined by $\{O_{task}, I_{task}, V_{task}, f_{Ctask}, f_{Ttask}, f_{Rtask}, f_{Atask}\}$

The task domain model is specific to the corpus on which the task is done. The techniques that are presented in section 4.2.1 enable the extraction of instances.

3.2.2 Example: science monitoring task

Science monitoring is an activity in which an expert observes and analyses the scientific, technical and technological environment of a domain in order to deduce, for example, the development opportunities or to analyze competition. Science monitoring consists in analyzing the evolution of the research themes studied and the different actors in the field.

An ontology referring to science monitoring activities has been developed manually in cooperation with the astronomers using Methontology (Fernandez 1997) based on work previously presented in (Mothe 2002). This ontology aims at proposing a cartography of the domain by presenting its actors: researchers, authors of articles, laboratories, countries and articles. Other knowledge is included such as temporal aspects, the literature of the domain in which scientific articles have been published, the subjects or themes of articles and the domain interests of the researchers. The last two concepts are particular concepts that make the link between the two ontologies (see section 3.4).

The associated science monitoring domain model is built semi-automatically from the corpus meta-data by populating the task ontology. The concepts and relations are chosen a priori, but automatic techniques extract their instances and values. These techniques are described in section 4.1.2.

The structure of the task model is defined by $\{I_{task}, C_{task}, R_{task}, A_{task}, T_{task}, V_{task}, CAR_{Rtask}, \leq^{C}_{task}, \sigma_{Rtask}, \sigma_{CARRtask}, \sigma_{Atask}, f_{Ctask}, f_{Ttask}, f_{Rtask}, f_{Atask}\}$ with

- C_{task} ={Researcher, Laboratory, Article, Country, Litterature, Date, Object, Journal, Book, Proceedings}
- R_{task} = {co-writes, works_for, has_as_domain_interest, writes, written_in, deals_with, published_in, is_situated_in}
- A_{task} = {has_firstname, has_surname, has_reference, has_address}
- I_{task}={ $researcher_1$,…., $researcher_{1n}$, $laboratory_1$,…, $laboratory_n$, $article_1$,…, $article_n$, $country_1$,…,$country_n$, $date_1$,…,$date_n$, $object_1$,…, $object_n$, $journal_1$, …, $journal_n$, … }
- T_{task} ={String }
- CAR_{Rtask}={symmetry, transitivity, functionality}
- V_{task} ={Dupond, Jean, A-1, A-2, 118 route de Narbonne 31400 Toulouse…}
- $\leq^{C}_{task}$ ={(Journal, Literature),(Book, Literature), (Proceedings, Literature)}

- σ_{Rtask} ={ (co-writes (Researcher, Researcher)), (works_for (Researcher , Laboratory)), writes(Researcher, Article), written_in(Article, Date), Has_as_domain_interest(Researcher, Object), deals_with(Article, Object), published_in(Article, Literature), is_situated_in(Laboratory, Country)}
- $\sigma_{CARRtask}$={(co-writes,symmetric)}
- σ_{Atask}={(has_firstname(Researcher, String),(has_lastname(Chercheur,String)), (has_reference(Article, String)), (has_address(Laboratory, String))}
- f_{Ctask}={(researcher1, Researcher), (laboratory1, Laboratory), …}
- f_{Ttask}={(Dupond, String), (Jean, String), (A-1, String), …}
- f_{Rtask}={(co-writes (researcher1, researcher2)), (works_for (researcher1, laboratory1)), (writes(researcher1, article1)), …}
- f_{Atask}={(has_surname(researcher1, Dupond)), (has_firstname(researcher1, Jean)), (has_reference(article1, A-1)), …}

The lexicon L_{task} :={L^{C}_{task},L^{R}_{task},F_{task},G_{task}}

- L^{C}_{task},={« researcher », « laboratory », « research institute », …}
- L^{R}_{task} ={« works for », « writes », « is situated in », « has as domain interest », …}
- F_{task} ={(Researcher, « researcher »), (Laboratory, « laboratory »), (Laboratory, research institute »), …}
- G_{task} ={(works_for, « works for ») (is_situated_in, « is situated in »), …}

A scheme of the ontology is presented in figure1.

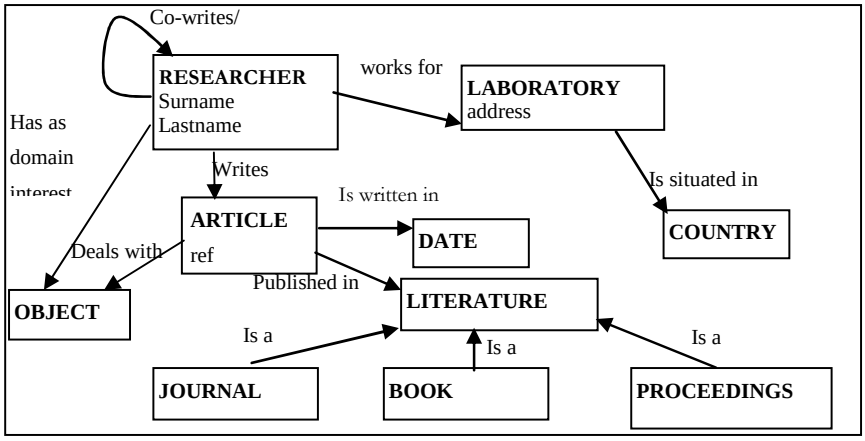


Figure 1 Science monitoring task ontology

3.3 Theme ontology

3.3.1 Definition

The theme ontology represents the domain knowledge treated in the content of corpus granules. It is represented through its lexicon and structures the concepts and their semantic relations in the domain. It aims at representing a consensual view on the domain in which the knowledge treated in the corpus can be situated. By accessing and browsing this ontology, the user will be able to situate his expertise and specify his information need according to his interpretation of the context.

The ontology is composed of

- A structure $S_{theme} := \{C_{theme}, R_{theme}, A_{theme}, T_{theme}, CAR_{Rtheme}, \leq^{C}_{theme}, \sigma_{Rtheme}, \sigma_{Atheme}, \sigma_{CARRtheme}\}$
- A lexicon $L_{theme} := \{L^{C}_{theme}, L^{R}_{theme}, F_{theme}, G_{theme}\}$

The theme domain can be completed with the domain model by populating the ontology with instances. The theme model is defined by the theme ontology previously described and by instances of its concepts that can be extracted from the documents and by the functions that instantiate these elements. The model is thus defined by $\{O_{theme}, I_{theme}, V_{them}, f_{Cthem}, f_{Tthem}, f_{Rthem}, f_{Athem}\}$

The theme model is generic and can integrate any ontology representing the domain treated in the corpus. As mentioned above, many domain ontologies are now freely accessible and can be reused even if elaborated for other purposes. However, the knowledge stated in the ontology must be in adequation with the corpus (Hernandez 2004). If a domain ontology does not exist or if the existing ontology does not represent the knowledge stated in the corpus, several methodologies have been developed in order to help build such representations from texts (Staab 2004). Methods have also been proposed to transform lexical resources such as thesauri into domain ontologies (Soergel 2004) (Hernandez 2005).

For practical reasons, the « theme domain ontology » is called « theme ontology ».

3.3.2 Example: astronomy domain case

Our case study was the application of our model to documents dealing with astronomy (astronomical objects, techniques, phenomena, theories). As no lightweight ontologies representing the themes of the corpus (international journal A&A[2]) existed, we proposed a semi-automatic method for transforming an existing astronomy thesaurus (IAU[3]) into a lightweight ontology by extracting recent domain knowledge from a reference corpus (Hernandez 2005). This method automatically extracts explicit knowledge from the thesaurus in order to define concepts with their lexicalization. It is also based on text analysis to extract new terms and meaningful semantic relations (“is a characteristic”,

[2] http://www.edpsciences.org/journal/index.cfm?edpsname=aa

[3] http://msowww.anu.edu.au/library/thesaurus/

"induces", "is measured by",…) which were not stated in the thesaurus. The resulting ontology contains 2547 concepts, each defined with 1 to 6 labels. Associative relations between concepts are of the type "is a part of", "is a property of", "is a phenomenon linked to", "is influenced by", "is determined by". Most of the work is done automatically. Manual intervention took 50 hours (mainly to validate abstract concepts needed to structure the whole ontology). The evaluation by experts showed that the automatically proposed concepts and relations were globally validated (with an average of 80% on the different aspects) and that they were helpful compared to a manual process. Part of the ontology, is presented in figure 2. Concepts are represented by rectangles; all the terms or labels referencing the concept are stated within the box (the main term is in bold).

As the context model of the theme domain interesting the user is situated more at the conceptual level than at the instance level, instances are not integrated in our model. Different tools (such as Simbad[4] integrated in a bibliographical server, ADS[5]) have been developed in order to access different specific astronomical objects, phenomena and theories present in a corpus. Our approach is different in that it presents a conceptual level. Taking instances into account, however, could be an extension to our work.

The structure of the astronomy theme ontology sample presented in figure 2 is $\{C_{theme}, R_{theme}, \leq^{C}_{theme}, \sigma_{Rtheme}, f_{Rtheme}, \}$ with

- C_{theme} ={Celestial_body, Asteroid, Comet, solar_system, Star, Sun, Solar_eclipse, solar_corona, astronomical_object}
- R_{theme} = {is_part_of, is_an_event_linked_to }
- $\leq^{C}_{theme}$ ={ (celestial_body, astronomical_object), (Comet, celestial_body), (Asteroid, celestial_body), (solar_system, celestial_body), (Sun, star)}
- σ_{Rtheme} ={ (is_a_part_of (Comet, solar_system)), (is_a_part_of (Asteroid, solar_system)), is_a_part_of (sun, solar_system), is_a_part_of (solar_corona, sun), is_a_event_linked_to (Eclipse_solaire, Sun) }

The lexicon L_{theme} :={L^{C}_{theme},L^{R}_{theme}, F_{theme},G_{theme}}

- L^{C}_{theme}={«celestial body », « comet », « asteroid », « planetoid », « solar system », « star », « sun », « solar eclipse », « solar corona » …}
- L^{R}_{theme} ={« is part of », « is an event linked to »}
- F_{theme} ={(Asteroid, « asteroid »), (Asteroid, « planetoid »), (Celestial_body, « celestial body »), … }
- G_{theme} ={(is_part_of, « is a part of ») (is_an_event_linked_to, « is an event linked to, »), …}

[4] http://simbad.u-strasbg.fr/Simbad

[5] http://cdsads.u-strasbg.fr/

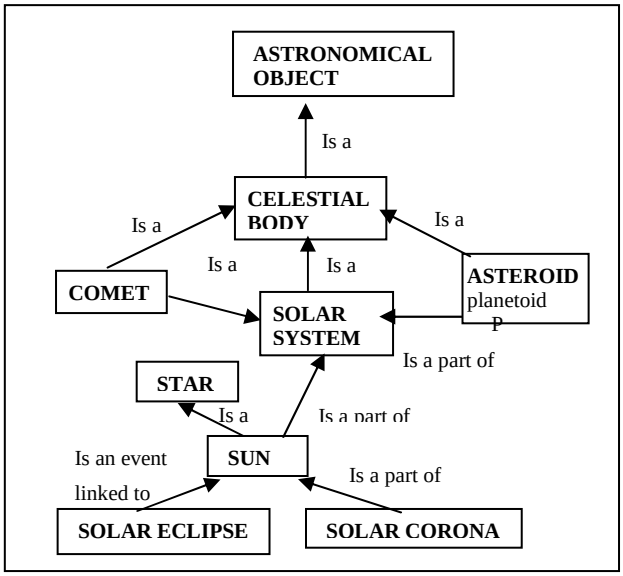


Figure 2 Sample of the astronomy theme ontology

3.4 Links between the two ontologies

3.4.1 Principle

The two ontologies are necessary to perform a search task and they must therefore be mutually accessible. The theme concepts can then be mapped to task ontology concepts, thereby situating them in the context of the task. The objects of the theme introduced in the task ontology can cover the whole theme domain.

The link is established through the concept *Object* belonging to the task ontology. It corresponds to the objects belonging both to the task domain and the theme domain. Its instances or values are taken from the theme model. The task model then defines the role of the theme model within the search.

Let the structure of the task domain model and the structure of the theme ontology be :

$S_{task} = \{I_{task}, C_{task}, R_{task}, A_{task}, T_{task}, V_{task}, \leq^{C}_{task}, \sigma_{Rtask}, \sigma_{Atask}, f_{Ctask}, f_{Ttask}, f_{Rtask}, f_{Atask}\}$ and $S_{theme} = \{I_{Theme,} C_{theme}, R_{theme}, A_{theme}, T_{theme}, V_{theme}, \leq^{C}_{theme}, \sigma_{Rtheme}, \sigma A_{theme,} f_{Ctheme}, f_{Ttheme}, f_{Rtheme}, f_{Atheme}\}$.

The *common_Object* belongs to C_{task}.

The instantiation or mapping function associated to the concepts belonging to f_{Ctask} is : f_c(common_Object,i), such as $i \in C_{theme} \cap I_{theme}$

Note the common objects between the two ontologies can be either an instance or a concept depending on whether the task focuses on instances of the theme domain (such as finding the *researchers having as domain interest binary star*) or on instances (such as finding the *researchers having as domain interest* a *specific binary star*).

3.4.2 Example: science monitoring in astronomy case study

The link between the science monitoring task ontology and the astronomical theme ontology in our case study is represented in figure 3. The link is established between an article and its themes

identified in the astronomy field. The instances are determined according to the most representative concepts of the granules (see section 4.1.2). Through the concept *Object*, the researcher's domain interests are also represented. The link is established automatically by the analysis of his publications.

The science monitoring ontology makes it possible to store, from the articles, the instances of the astronomy domain that are treated in the corpus. The subjects are stored as instances of the concept objects.

For science monitoring purposes, the links established between the two ontologies makes a new analysis possible. The themes of the research organizations can be analyzed according to the domain interest of their researchers, the evolution of the themes can be observed over time, and the impact of a researcher on the laboratory's themes from the date he joins the lab.

Examples of links are presented in figure 3. The researcher whose surname is *Dupond* and first name is *Jean* has as domain interest the concept *solar corona* from the theme domain, the *article* that has the reference *A-1 deals with* the theme domain concepts *solar eclipse*. Both concepts correspond to instances of the concept "object" from the task ontology.

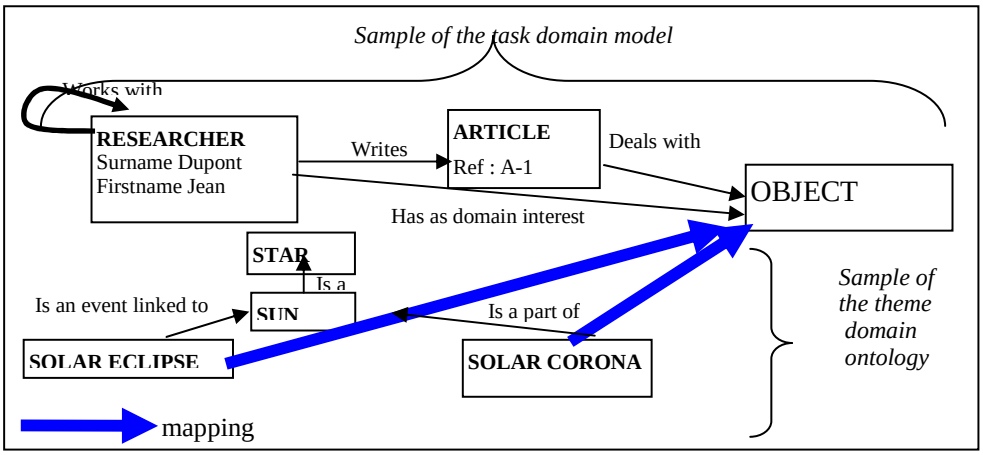


Figure 3. Links between the task domain ontology and corpus domain ontology

The links between the two ontologies are then formalized by

$f_{Ctask} = f_{Ctask} \cup \{(common\_object1, solar\_eclipse), (common\_object\_2, solar\_corona)\}$

$f_R = f_R \cup \{ deals\_with (article1, solar\_eclipse), has\_as\_domain\_interest (researcher1, solar\_corona) \}$

The context, represented by both the corpus theme ontology and the task ontology, is integrated in the IR process in two ways. Firstly, the granules are situated in this context when indexed according to the two ontologies. Secondly, the two ontologies are used as supports for accessing the information. These two aspects are developed in sections 4 and 5.

4 Indexing of documentary granules

The two ontologies representing the search context are used to index the granules. Indexing plays an essential role in IRS by defining the descriptors which represent the granules and through

which the granules can be accessed and analyzed. Two approaches, one for each type of ontology used by the model, are adopted to index the granules.

The first consists in semantically indexing the granule contents through the theme ontology. As explained previously, this type of indexing is the extension, in IR, of the use of terminological resources such as concept hierarchies and thesauri by adding more formal semantics. The aim is to choose descriptors of the granule content through the objects of the ontology referenced rather than the words present (Haav 2001). Semantic indexing is only possible through the existence and use of resources that explicitly describe the information corresponding to the objects. Our approach is original in that this indexing not only uses the hierarchical links but also the associative links described in the ontology. The dependence of concepts in the ontology and the granule is thus taken into account. As the granules may refer to different concepts, a concept indexing weighting function is introduced which, for each granule, returns the concepts in function of their representivity of the content of the granules. Semantic indexing according to the theme ontology is done automatically.

The second approach derives from the Semantic Web. In this context, indexing (also called granule annotation) has a double objective: to index the content of granules and to represent granules by generating corresponding meta-data. This approach is based on ontologies that model objects of the world through the actors and entities that the granules constitute and include (Guha 2003). In our case, granule annotation is done through the task ontology. A manual mapping is made between the document meta-data and the task ontology concepts. However, more sophisticated techniques such as Name entity extraction presented in (Kiryakov 2004) could be easily integrated.

The mechanisms implemented for the two types of indexing are described in the following sections.

4.1 Semantic indexing of granules through the theme ontology

To determine the themes dealt with in the corpus, a semantic granule indexing mechanism using the theme ontology is used. It consists in searching for the granule content among the concepts of the ontology. The granules are then indexed, not by terms that may be ambiguous but by concepts for which the semantics are clearly defined.

The semantic indexing process we propose comprises two stages: identifying concepts in the granules and weighting them in order to determine the representativity of the granules. The method is novel in that the concepts are weighted according to their organization in the structure of the lightweight ontology, in addition to their frequency of occurrence in the granule and the corpus. Contrary to existing approaches (Baziz 2005) (Desmontils 2002), the weighting takes into account all the relations of the ontology's structure and not only taxonomic relations. When presenting the process, the different indexing stages are illustrated with examples from the astronomy field on which it has been evaluated.

4.1.1 Identification of concepts in granules

To identify concepts referenced in the textual granule contents, the lexicon of each granule is first extracted. A syntactic analyzer (Syntex [6](Bourigault 2000)) is used to extract the syntactic phrases and nouns of each granule. Syntactic rather than statistical extraction is preferred as concept labels are generally composed of phrases and precision is essential when extracting them from granules. The advantage of these techniques is that phrases are detected according to the grammatical roles of the words in the sentence and not only according to co-occurrence (as in statistical techniques). Nominal phrases only are considered as they correspond to the nature of the concept labels in the ontology. Syntactic phrases are extracted from a granule in different forms: the maximum form corresponds to the phrase composed of all the expansions linked to the main term of the phrase; the reduced forms correspond to the different phrases derived from the maximum phrase from which the expansions have been successively eliminated. The advantage of considering the different forms into which a phrase may be decomposed is that the concept labels in the ontology may be searched for in these forms. For example, in the sentence « magnetic connection between black holes and disks are observed », the maximum phrase « magnetic connection between black holes and disks » is extracted. One of the main terms is « connection »; it will give rise to the reduced phrase « magnetic connection ». Other reduced phrases will be « black hole » and « disk ». If the maximum phrase does not appear in the lexicon of the ontology, the reduced phrases may be present. It must be noted that to capture the lexical variations of the phrases, syntactic phrases are extracted in their lemmatized form. Finally, the lexicon of a granule is defined by all the lemmatized nominal phrases extracted in their various forms.

The lexicon of a granule is noted by:

$$\mathbf{L_{granule} = \{lemmatized\ nominal\ phrases\}}$$

The identification of the concepts referenced in a granule consists in finding, among the phrases that make up the lexicon of the granules, the labels of the ontology and the corresponding concepts. These two steps are described in the following sections.

Identification of labels

The identification of labels consists in finding, in the lexicon of the granule, the phrases corresponding to the labels of the ontology. This implies evaluating:

$$\mathbf{L_C \cup L_{granule}}$$

The labels referring to the most specific concepts can thereby be identified. The labels of the ontology are in fact looked for in the set of phrases of each sentence of the granule. The phrases extracted for each sentence include its maximum and reduced phrases. As in (Vallet 2005), to ensure that the most specific labels are identified, the labels that are chosen are those which correspond to the phrases in their longest form (i.e. composed of the greatest number of words).

[6] We chose this system because it has been developed locally and shown to be efficient but any syntactic analyser could be used.

For example, for the sentence « *magnetic connection between black holes and disks are observed* » belonging to a granule, the phrases : « *magnetic connection between black hole and disk* », « *magnetic connection* » « *connection* », « *black hole* », « *hole* » and « *disk* » are extracted. If « *magnetic connection* », « *connection* », « *black hole* » and « *hole* » are labels of concepts in the ontology, then the most specific labels in the sentence of the granule will be « *black hole* » and « *magnetic connection* ».

Identification of concepts

This step consists in identifying for each label the associated concept:

$$c/\ F(l)=c \text{ for } l \in L_C \cap L_{granule}$$

When the label found in the granule corresponds to several concepts, a disambiguization mechanism is used to determine the concept actually referred to in the granule. First, the context of the occurrence of the label in the sentence is examined using other non-ambiguous concepts found. If no non-ambiguous concept is found in the sentence, the label is analyzed in relation to the concepts found in the granule. The concept chosen is the one that is semantically closest to the other concepts identified in the documentary context considered. In order to identify the semantic link between the various candidate concepts and the context concepts, the distance between the concepts is evaluated according to the number of relations with the shortest path separating the concepts in the ontology. For each candidate concept, the sum of the number of relations is calculated. The concept proposed is the one for which the distance is shortest.

Consider for example, the following sentence extracted from a granule: « ***Polarization*** varies noticeably with emergent *photon* energy below 40keV, being up to 30% and down to -10% for different angles of view; these variations cover the range of observed *magnitudes*.» and the sample of the astronomy domain ontology presented in figure 4 (concepts are represented by boxes in which the labels are stated, relations are represented by arrows above which the label of the relation is given).

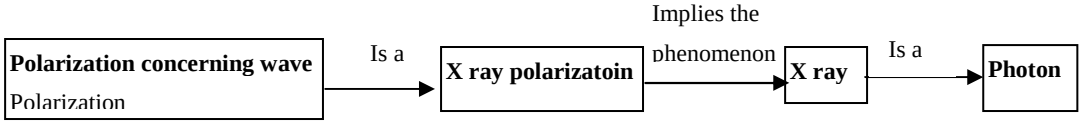


Figure 4 Sample of the astronomy domain ontology

Different concepts can be referenced with the label « *polarization* »: « *polarization concerning wave* », « *polarization concerning charge separation* ». The non-ambiguous labels « *photon* » and « *magnitude* » are extracted from this sentence; they reference the concepts « *photon* » and « *magnitude* ». To identify the concept referenced by the ambiguous label « *polarization* », the number of relations separating the two concepts that it references (« *polarization concerning wave* », « *polarization concerning charge separation* ») and the other concepts identified in the sentence (« *photon* » and

« *magnitude* ») is calculated. The number of relations separating « *polarization concerning wave* » and « *photon* » is 3 since « *polarization concerning wave* » is the father of « *X Ray polarization* », itself linked by the relation « *is a phenomenon linked to* » to « *X Ray* », itself sub-child of the concept « *photon* ». The number of relations separating « *polarization concerning wave* » and « *magnitude* » is 7. The sum of the relations separating the concept « *polarization concerning wave* » and the concepts identified in the documentary context is therefore 10. The same calculation is made for the concept « *polarization concerning charge separation* » and the sum is 30. The concept chosen to reference the label « *polarization* » in the sentence is therefore the concept « *polarization concerning wave* » since its semantic link to the other concepts is strongest.

At the end of this step, the concepts referenced in all the granules are identified.

4.1.2 Weighting of concepts

The weighting of a concept determines the degree to which a concept in an ontology is representative of a given granule, but it also reflects its discriminatory power, i.e. its capacity to distinguish between relevant and non-relevant granules when this concept is considered. This degree is called the representativity of the concept in the granule. As the concepts are not independent in their use either in the granules or in the ontology, the degree of semantic representativity must take this dependency into account. We thus propose a measure for weighting concepts that includes both the statistical and semantic representativity of the concept in the granule.

The **statistical representativity** is calculated by adapting the measure tf.idf (Roberston 1976) used in IR to calculate the discriminatory power of a term. Applied to concepts, this measure favors concepts with a high frequency in the granule but a low frequency in the rest of the collection.

The formula proposed is:

$$\text{Re}\,presentativity\_stat(c,g)=cf_{c,g}\times idfc \quad \textbf{(1)}$$

$$idf_c=\log(\frac{N}{f_c})+1$$

where $cf_{c,d}$ represents the frequency of occurrence of the labels of concept c in granule g

and f_c corresponds to the number of granules containing at least one of the labels of concept c

Semantic representativity takes into account the link in the ontology between the concept considered and the other concepts of the granule and illustrates the semantic context of the concept. It is based on the principle that the semantically closer a concept is to the other concepts found, the more representative it is of the themes of the granule. This semantic representativity is calculated by the semantic proximity of the considered concept to the other concepts found in the granule and is calculated by formula (2).

$$\text{Re}\,presentativity\_sem(c,g)=\sum_{ci\in\{\forall cj\in GC(g),c\neq cj\}} prox(c,ci) \quad \textbf{(2)}$$

where GC(g) represents all the concepts found in granule g.

The proximity measure prox is presented in the section 4.1.3.

In order to combine both **concept representativity** factors, a concept is weighted by the sum of its statistical and semantic representativity, each having been normalized on the whole of the collection. The factors α and β allow them to play a non-symmetric role. Their impact has not yet been studied.

$$\mathrm{Re}\,presentativity(c,d)=\alpha\frac{\mathrm{Re}\,presentativity\_stat(c,d)}{\max_{i,j}\left(\mathrm{Re}\,presentativity\_stat(cj,dj)\right)}+\beta\frac{\mathrm{Re}\,presentativity\_sem(c,d)}{\max_{i,j}\left(\mathrm{Re}\,presentativity\_sem(cj,dj)\right)} \quad \textbf{(3)}$$

These formulas are generic as a granule may correspond to a document, part of a granule, a set of documents. It is therefore possible to calculate the representativity of a concept for an extract or for a group of documents (even a complete corpus).

4.1.3 Measure of proximity between concepts

In order to evaluate the proximity of concepts, the structure of the ontology and the information contained in the corpus on the concepts must be taken into account. The measures in the literature that are most suitable are those that are based on the information content of concepts (Resnik 1995) (Jiang 1997). In these measures, the information carried by the concepts is captured by the probability of obtaining the concepts in a corpus. The proximity or semantic similarity of the concepts is then evaluated by their common information. These measures to define similarity between concepts are defined in a taxonomy. The common information is reflected in the information content of the most specific concept generalizing the two concepts. These measures apply to the evaluation of the proximity of concepts chosen for document indexing since they evaluate the semantic links using the hierarchical structure of the ontology and the information contained in the corpus concepts. Unlike the measure used in (Desmontils 2002) which gives the same weight to all « is a » relations, these measures evaluate the semantic proximity according to both the ontology « is a » relations and the corpus. The relations present in a lightweight ontology are, however, of several types (taxonomic, meronymic, causal, transitive or with no logical property). Not to take these associative relations into account implies ignoring part of the semantics contained in the corpus. We propose therefore to extend these measures based on the information content of concepts and to include these relations.

The measures based on the concept information content evaluate the similarity of two concepts according to the information content of the concept generalizing them obtained by considering the most specific concept to which the concepts are linked by the relation « is a ». To extend these measures to non-taxonomic relations, the first solution proposed in (Lord 2003) is to consider the associative relations as « father-son » relations, in a way, transforming them into « is a » relations. However, this solution implies considering a concept as generalizing a given concept even though it has no semantic link with the concept in question. To illustrate this, in the example presented in figure 5, we have chosen an intuitive example not linked to the astronomical domain. The associative relations (« *belongs to* », « *possesses a* ») are transformed into « father-son » relations. Intuitively, it appears wrong to consider that the concept « Address » generalizes the concepts « Car »

and « Shop ». The semantic link between the concepts considered and the generalizing concept is retained in the initial formulas by the transitivity of the relation « is a ».

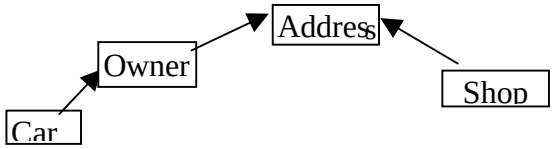


Figure 5 Example of associative relations transformed into « father/son » relations

We propose to adapt the measure defined by Jiang (Jiang 1997). The choice of this measure was motivated by the intuition underlying this measure that relations do not represent the same semantic weight according to their nature (taxonomic relations, non taxonomic relations …). The semantic distance represented by the relation «is a » appears intuitively greater than the distance between two concepts linked by another type of relation. The measure proposed by Jiang aims at quantifying the semantic distance between two concepts from the weight of the semantic relations separating the concepts by the shortest path. Although this measure differentiates the weight of relations, only the weighting of taxonomic relations is proposed. Formulas (4) (5) (6) and (7) calculate this distance.

Dist(c1,c2)= where pcc(c1,c2) represents all the concepts with the shortest path

father(c) is the parent concept p of c by a taxonomic relation

$$Weight(c,father(c))=CI(c)-CI(p) \quad \textbf{(4)}$$

$$IC(c) = -\log (p(c)) \quad \textbf{(5)}$$

$$Freq(c)=\sum_{n \in word(c)} count(n) \quad \textbf{(6)} \qquad p(c)=\frac{freq(c)}{N} \quad \textbf{(7)}$$

where word(c) represents all the terms or labels representing concept c and the concepts subsumed by c, count(n) the number of occurrences of word n in the corpus and N the total number of occurrences of the labels of concepts found in the corpus.

We propose a new measure inspired by the information content approaches.

Valid path

The solution we propose is to consider the semantic link between two concepts only from the paths deemed valid by the type of their characteristic (transitivity, symmetry ...) when specified in the formalization on the ontology. We consider a path valid if it does not involve more than one change of direction. A change of direction is represented by a non-transitive property. For example, when considering a transitive relation R and the following facts aRb and bRc, it is possible to infer that aRc, and that a and c are linked by the same property. However, when a relation is non transitive, this type of inference cannot be made and the relation implies a change of direction between concept a and concept c. Taking into account changes of direction is important for a measure based on the information content of concepts as it guarantees that the concepts considered have common

characteristics. When the link is symmetric, the relation is considered in both directions. The path involving the two concepts is considered valid in both directions.

Measure

In the case of the relation « is a », the information content of the concepts is calculated using their descendants resulting from « is a » relations as proposed originally in (Resnik 1995) and applied in (Jiang 1997). In our case, the frequency of a concept is calculated after its disambiguization. Unlike in the formula proposed in (Jiang 1997), the information content of a concept is not biased by the occurrence of labels that may reference another concept.

In the case of non-taxonomic relations, the weight of the link cannot be evaluated by formula (4) proposed by Jiang as the information contents of the two concepts have no common elements. The information content is not captured by any common label. We propose therefore to evaluate the weight of these relations using a new element called information content of the relation. This is calculated by the probability of co-occurrence of the two concepts in the corpus. This probability is established by the number of times the labels of concept c1 and the labels of its descendants co-occur with labels of concept c2 and its descendants. As for the information content of concepts, the function –log is used to reduce the information content with a high probability link. Formulas (8) and (9) calculate the weight of a non-taxonomic link.

$$\text{Weight}_{\text{non\_taxo}}(c,c_2)=\text{CI}(c,c_2) \quad \textbf{(8)}$$

$$\text{with CI}(c,c2)=-\log\left(\frac{nbdoc\_cooc(c,c2)}{nbdoc\_total}\right) \quad \textbf{(9)}$$

where nbdoc_cooc(c,c2) is the number of documents in which c and c2 or each of their descendants co-occur and nbdoc_total is the total number of documents in the corpus

The formula proposed to calculate the distance between two concepts c1 and c2 is as follows:

$$\text{dist}_{\text{prop}}(c1,c2)=\sum_{c\in\, pccvalide(c1,c2)} Weight_{prop}(c,rel(c)) \quad \textbf{(10)}$$

where pccvalide(c1 and c2) is the shortest valid path between c1 and c2

and rel(c) is the concept linked to c in the ontology

$$\text{Weight}_{\text{prop}}(c,c_2)=\text{CI}(c)-\text{CI}(c_2) \text{ if } c_2=\text{father}(c) \quad \textbf{(11)}$$

$$\text{Weight}_{\text{prop}}(c,c_2)=\text{Weight}_{\text{non\_taxo}}(c,c_2)=\text{CI}(c,c_2) \text{ else}$$

The proximity is then calculated from the inverse of the distance with formula (12).

$$\text{Prox}_{2\_\text{prop}}(c1,c2)=\frac{1}{1+dist_{prop}(c1,c2)} \quad \textbf{(12)}$$

4.2 Annotation of documents through a task ontology

The annotation of granules by the task ontology consists in searching in the granules the values of the meta-data represented in this ontology. If the meta-data values are explicitly present in the granules, the techniques implemented are those that are presented in section 4.2.1. If the values are extracted through the corpus theme ontology (thus corresponding to common objects of the two ontologies), the techniques used are those described in 4.2.2.

4.2.1 Extraction of instances occurring in the granules

Knowledge extraction techniques are used to extract instances of concepts corresponding to meta-data explicitly present in the granules. In our case, an extraction mechanism is introduced in the analysis of the granule tags. A manual correspondence is established between the tags and concepts corresponding to the meta-data, either with the DTD if the documents are represented in XML or with the particular format of representation of the corpus. For each of the character chains inside the tags, an instance of the corresponding concept is created. When the granules are not explicitly structured, more sophisticated information extraction techniques can be used (Dkaki 1997) (Amous 2001). When the granules come from a document (or part of a document), the meta-data are implicitly extracted by the meta-data associated with it.

4.2.2 Extraction of instances of an object common to the two ontologies

The instances of the object concept in the task ontology that are implicitly present in the granules are extracted thanks to the analysis of the granules containing information targeted by the object.

As the object takes its values from the theme ontology, the instances are determined by the representativity of the theme ontology concepts in the whole set of granules considered. The semantic indexing used was presented in part 4.1.

For example, in the case of the ontology for the scientific watch task presented in figure 3, the instances correspond to those of the concept *« object »*. When the instances sought concern a researcher's field of interest, the instances corresponding to the concept *« object »* are the concepts considered representative of his publications. These concepts are extracted in function of their representativity. The set of granules considered is no longer the corpus but the researcher's set of publications.

What is original in this approach is that the concepts extracted from the theme ontology are weighted in function of their representativity. When a granule or set of granules treats many concepts, the concepts are organized in function of this score. As these concepts may be instances of the task ontology, a weight is also associated with these instances, making it possible for an information access system to return instances in function of their relevance by targeting the returned information.

5 Access to information

An interesting aspect of our approach is that the context is taken into account to provide the user access to the information contained in the corpus. A prototype of an interface based on the model has been developed. The interface is developed to support OWL (McGuinness 2004) ontology visualization. A snapshot of this interface is presented in figure 6. With the interface it is possible to visualize both the task ontology and theme ontology. The exploration of the corpus is thus done through the knowledge related to both aspects of the context. For this purpose, the screen is divided into two windows, the left window showing the task ontology, the right window showing the theme ontology. The ontologies are represented by the concept labels in yellow rectangles, and by the relation labels above the arrows indicating the relations between two concepts. The interface contains different icons making it possible to open the two ontologies, load the corpus and zoom on certain parts of the ontologies.

As pointed out by (Thompton and Croft 1986), the user is better able to recognize what is wanted than to describe it. We have thus developed a browsing interface to provide two views on the available information. The first navigation level is situated at the concept level to give an overview of the content of the collection and associated knowledge. The interface provides the user with a second level of navigation at the instance level in order to go into the details of the collection contents. Traditional browsing interfaces (such as cat-a-cone for example) do not provide the user with such functions as the user can only browse the domain by going deeper into the hierarchy of concepts (no instances are extracted within the documents).

Moreover, at each navigation level, the user can choose to focus on regions of the ontologies. He can thus explore specific parts of the domain he is not familiar with or he wishes to investigate. Compared to existing systems, in our interface the semantic context of each object is specified according to all its relations defined in the domain. This enables the user to understand the meaning of each object according to his own knowledge.

Another interesting point is that the user can decide to access documents according to either the theme aspect of his need or to the specific data he is interested in. By browsing the task ontology he can have access to all the knowledge stated in the corpus for an object he is focusing on. With traditional IRS, the user has to make specific queries to access information on each facet of the object.

In order to illustrate the originality of our interface, the following paragraphs describe the specific needs of the astronomers that could not be satisfied with traditional IRS and that are met with our system.

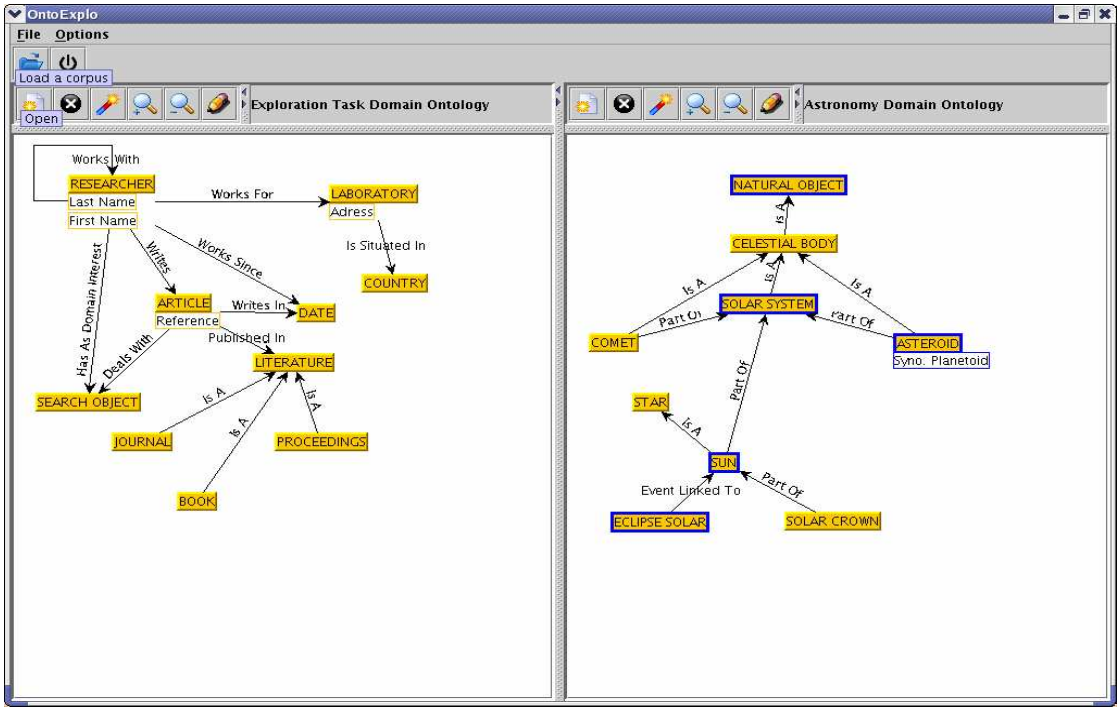


Figure 6 Corpus exploration interface

5.1 Users can decide on the focus

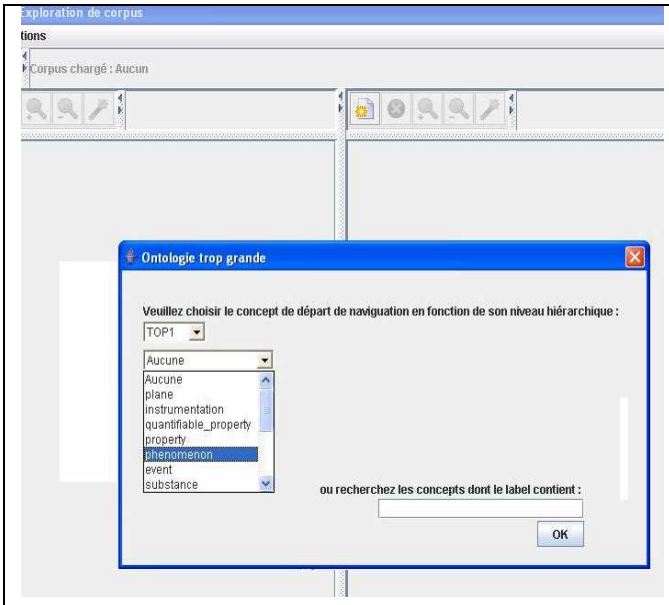
Astronomers consult the publication collections with different domain interests. As in any other domain, users know some of the terms and can formulate a query; but they do not necessarily know all the vocabulary and the different ways a concept can be expressed in a document. Moreover, most of the time they are specialists of a sub-field and their queries are restricted to specific sub-fields. They want the system to be adapted to these different contexts.

The fact that the vocabulary is formalized in an ontology solves the problem of different terms representing a concept. For example, in a traditional IR system documents dealing with “planetoid” and “asteroid” would be retrieved separately. With an ontology, they are associated to the same concept. Moreover, considering an ontology enables the system to distinguish polysemic terms. For example, when a user enters “polarization” the system indicates to the user that this term refers to several concepts (« *polarization concerning wave* », « *polarization concerning charge separation* ») that the user might not be familiar with.

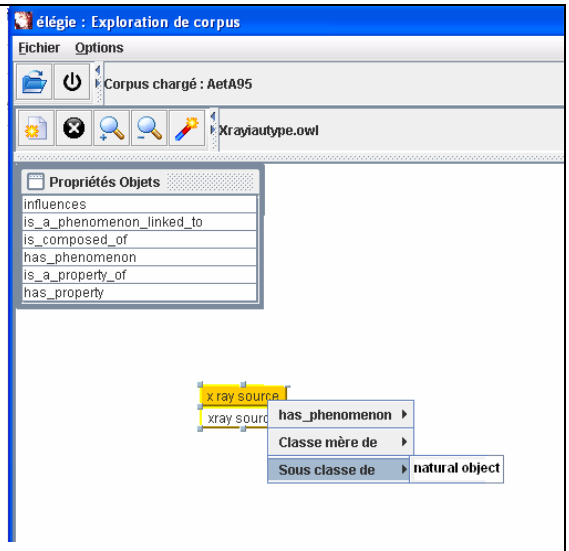
By providing access to the corpus through the browsing of concepts of the two ontologies, the a priori knowledge of each user can be confronted with the knowledge used to represent the context of the search. With this presentation the user is more likely to situate the different concepts according to his own knowledge and to interpret the semantic context of each of them through the relation to the concepts he knows.

More interestingly, the interface allows the user to focus on the part of the ontologies he wants to visualize. This is important since the ontology can be huge (in our case it contains more than 2500 concepts) and the user is generally interested in a restricted part to query the system. To do this, the user starts from an empty window, adds a concept he knows (typing one of its labels i.e. a term) and

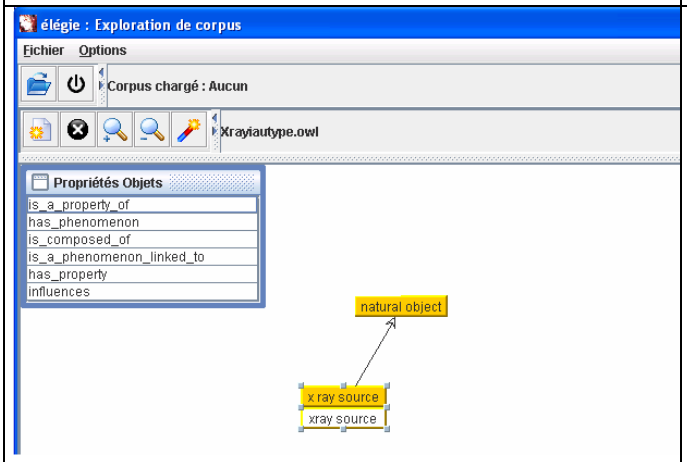
then recursively adds other concepts that are linked to the displayed concepts. To do that, he simply selects the starting concept and asks for related concepts in the ontology (he can choose the type of links he wants to visualize). This functionality is illustrated in figure 7.



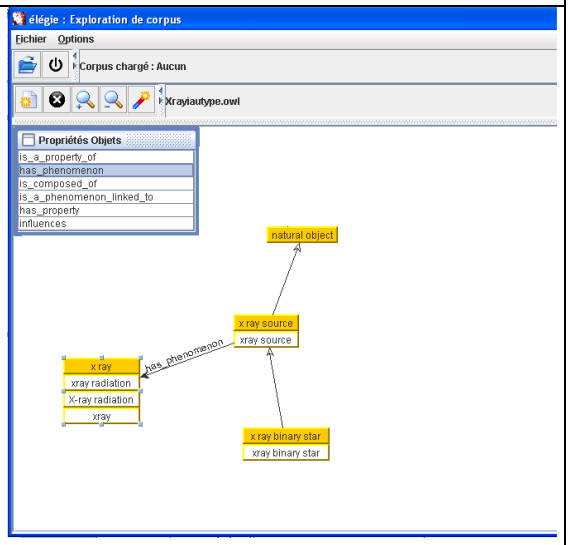
a) The user typed the term "x ray source" The corresponding concept (x_ray_source) is displayed



b) By selecting this starting concept, he accesses different types of relationships. He selects the relation "sub-concept".



c) Following action b), related concepts are displayed. The user can, step by step, browse some of the branches or links encoded in the ontology.



d) After some steps, the user visualizes the part of the ontology he is interested in.

Figure 7: viewing a local part of the ontology to focus on user's interest.

The user can however navigate the whole ontology if he prefers and even hide some parts.

5.2 User can decide on the abstract level

Astronomers are interested in global and specific views of the information available. This can be explained by the fact that sometimes they know the field and what they are looking for, other times they know less and first need a global view. Our system meets these two types of demands.

The first abstraction level provided is the concept level where instances are not shown. Regarding the task ontology, this means that the user does not visualize the names of the researchers, the references to the potentially interesting documents or other values of the meta-data associated with the documents. Regarding the theme ontology, it means that the users would not visualize the different instances of a “double star” for example, but only visualize the fact that a “double star” is a “star” and other types of semantic links between concepts.

By using the second level of access to information situated on the instances in the task ontology, the user can have details on the accessible content of the collection when he visualizes the objects of the collection and their links. For example, he can visualize the names of the authors in the document set that has been indexed. The user is guided by the target elements of his search and can specify his views on the domain.

Each level is accessible by the other. The choice of a concept may therefore lead to the visualization of its instances and, from an instance, it is possible to know the role of the associated concept in the task.

5.3 Exploration from the task ontology

The astronomers wanted to access specific data useful for a scientific watch task on the domain. They needed not only access to the data, but also to be able to analyze the co-relations and semantic links between them. With their previous system, they had to formulate specific queries to analyze each correlation (authors working together, domain interest of each researcher, publish articles of each author). By browsing the different instances of the task ontology, the user discovers all the actors and meta-data detected in the corpus (left-hand side of figure 8). For example, by selecting this instance (figure 8) which is the object of his query in the task ontology (e.g. Researcher Cecchini S.), the user visualizes the known relations between this instance and the other instances: researchers with whom he works (figure 9): Giaconelli M. …, articles: ref 124, 148, 789, laboratory (affiliation in his publications): TESRE. The link between the two ontologies is also presented to the user. The search object concept is colored in red in the task ontology (left side). The instances of the search object concept linked to Cecchini S. in the task ontology (field of interest of a researcher) are then found in the theme ontology and are presented to the user on the right side of the interface in the content domain ontology (figure 8). These concepts are also colored in red (the same color as the search object concept in the task domain ontology) in order to be distinguished from the other concepts of the theme ontology which are not his domain of interest. Cecchini’s centers of interest are solar systems and eclipses. The user can thus situate these concepts in their context. The concepts are in effect presented in samples of the theme ontology, thus making it possible to visualize the concepts to which they are related.

Astronomers also wanted to be able to access document content and interpret quickly the document themes. Through the interface, a user can select the reference of the article, either from the list of instances of the concept article (see figure 8), or from the instances of the articles written by a

specific author (see figure 9). As shown in figure 10, a pop up window containing the article appears and the information displayed in the window is automatically updated. The information known about this article is presented in the task ontology frame and all the themes of the domain treated in the article are presented in the theme ontology frame as instances of the search object concept. The user thus has information about the document. The theme 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.

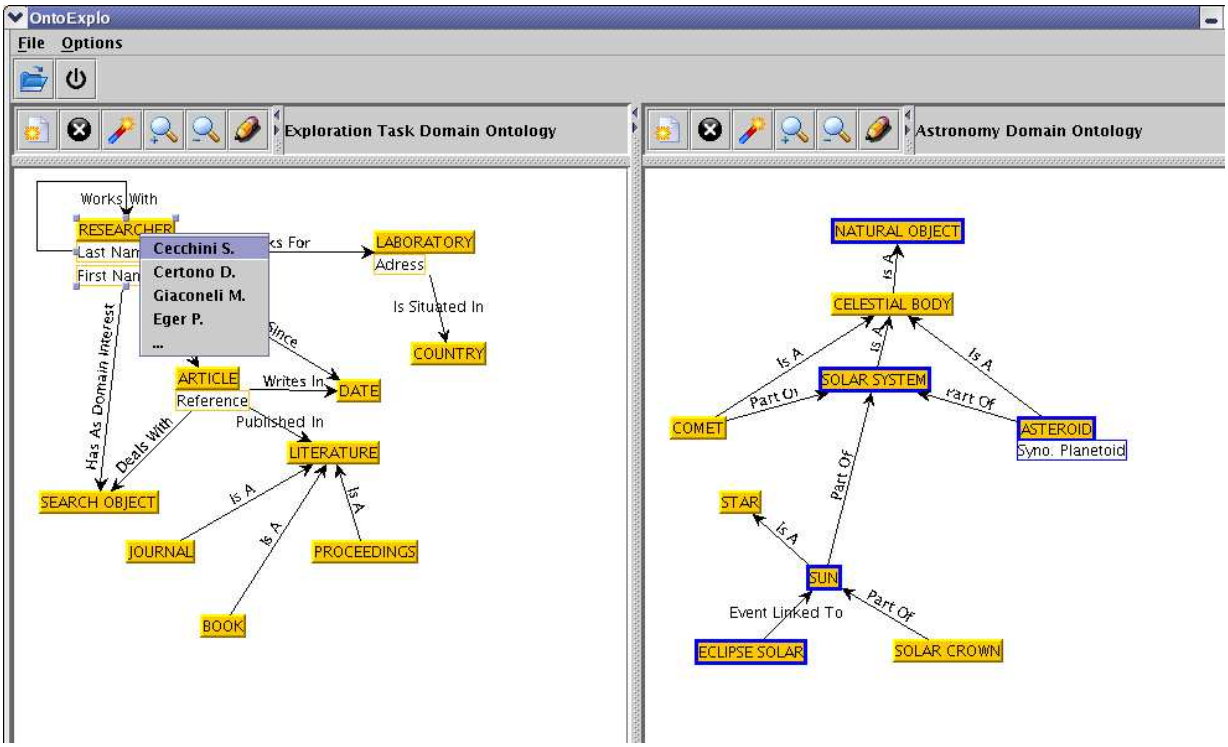


Figure 8. Visualization of the instances of the task ontology

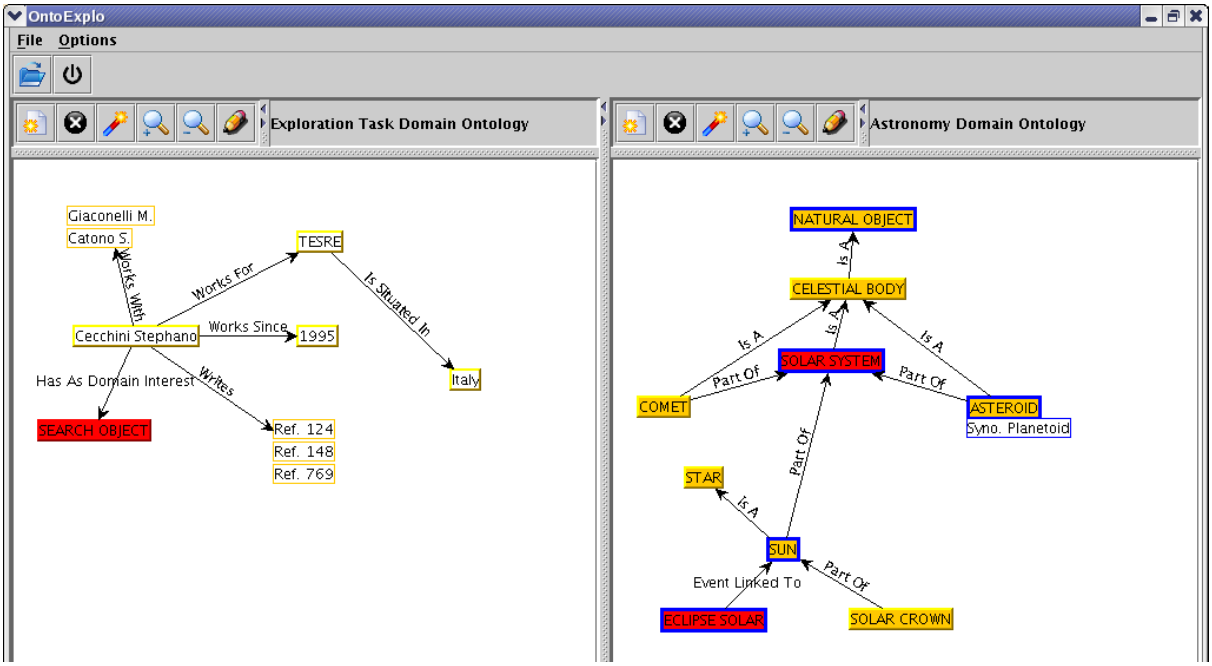


Figure 9. Visualization of the knowledge learnt for an instance researcher of the task ontology

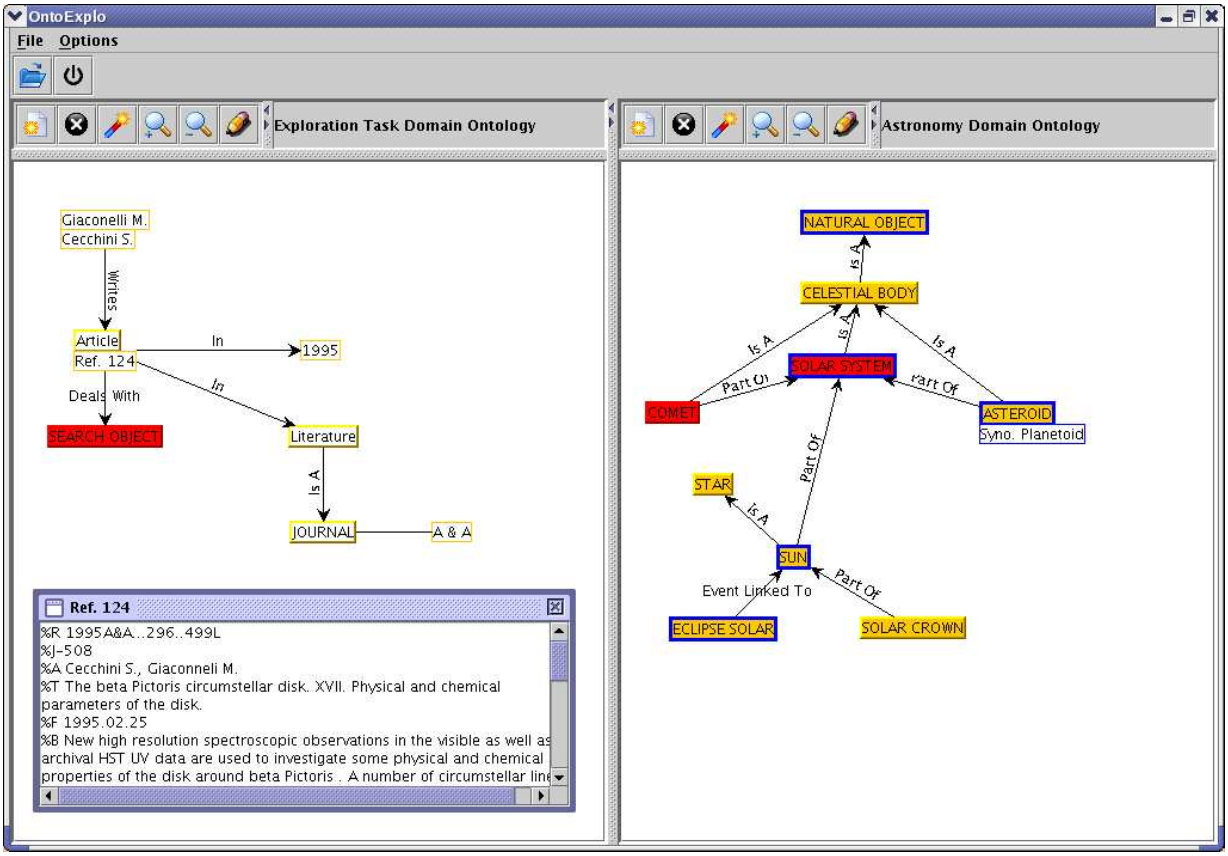


Figure 10. Visualization of the knowledge learnt for an instance Article of the task ontology

5.4 Exploration from the theme ontology

Apart from the specific data linked to the search task, astronomers wanted to access documents dealing with chosen themes of the domain. In order to provide such access, the interface

proposes to browse through the theme ontology. The user explores the corpus according to the knowledge related to the content of the documents. This kind of browsing enables the user to find information about a specific theme of the domain.

This possibility is illustrated in figure 11. All the concepts and relations of the ontology are shown. 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 (for example, selecting Cecchini S., figure 8 will be shown to the user). In the same way, when the user is interested in accessing the article dealing with this subject, he selects it (figure 8); the windows are updated: a pop up window containing the article appears and the information linked to this article is represented as in figure 9.

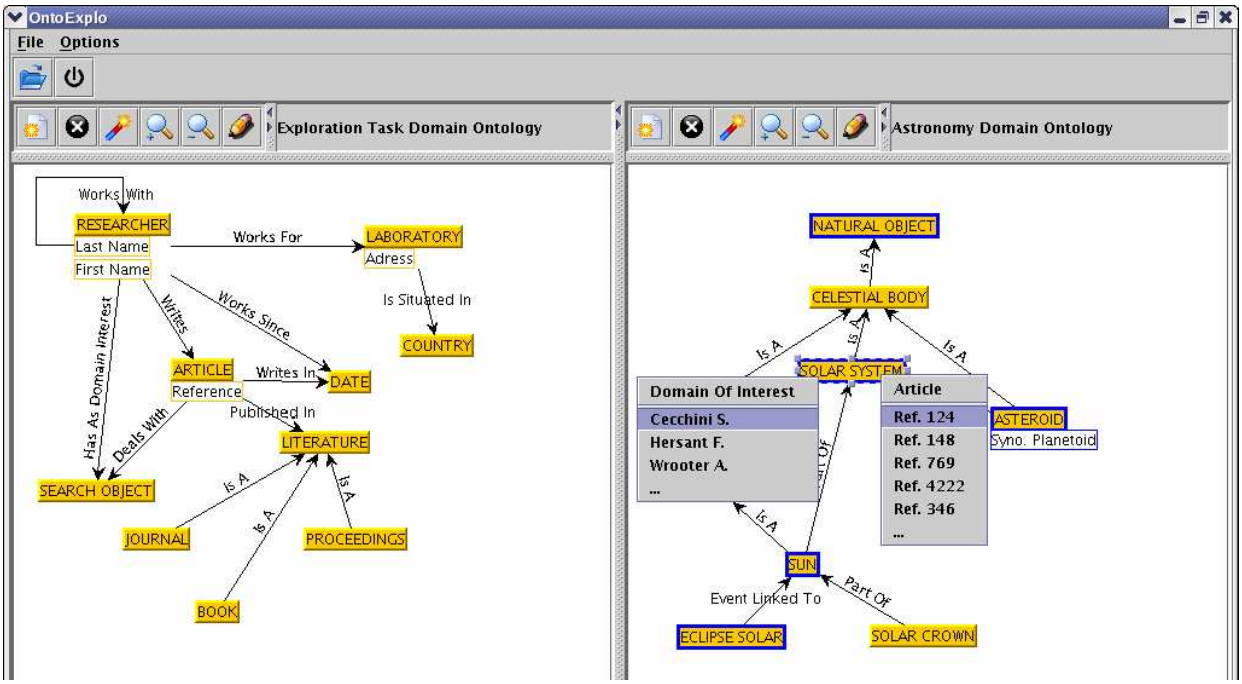


Figure 11. Visualization through the theme ontology

6 Experimentation

In order to evaluate our approach, we analysed separately the performance of the semantic proximity measure, the result of the indexing process and the access to information in the field of astronomy. In this field, huge amounts of data are produced daily, and very precise indexing techniques are necessary for managing and accessing them efficiently.

The document corpus considered is composed of article abstracts published in the journal Astronomy and Astrophysics (A&A) in 1995. The ontologies used were elaborated in the context of a project involving astronomers from a French observatory.

Two astronomers were asked to evaluate the different propositions in order to have a representative point of view.

The measure of proximity between concepts (section 4.1.3) was evaluated by its correlation to the astronomers' judgments of semantic proximity for 30 concepts extracted from the astronomy ontology. The results were compared to those obtained with two measures of the literature (Jiang 1997) (Lord 2003). Precise details are not given as this measure is not the central aim of the work presented. The correlation of our measure to human judgments was 0,47 compared to 0,11 and 0,32 for the two other measures. This can be explained by the two hypotheses that motivated our measure. First of all, that non-taxonomic relations have to be taken into account when calculating similarity. Secondly, the intuition by which two concepts linked by several non-taxonomic relations do not carry common information confirmed by the fact that our results are better than those obtained by Lord's measure.

The semantic indexing process (section 4.3) was evaluated on 100 publications in the astronomy domain to assess its overall precision. The concepts automatically extracted for each of the documents were compared to concepts manually extracted and ranked by two domain experts. The results showed that 80% of the extracted concepts were highly relevant and that their ranking corresponded for 90% to the manual ranking. Compared to terms extracted according to the tf.idf measure, the astronomers pointed out that the concepts were much more significant in the sense that they were not ambiguous and that their ranking was more representative of the general content of the documents.

Concerning information access, the two astronomers were asked to perform the different cases presented in section 5. No formal evaluation was made, only interviews with the experts. Both of them preferred the new solution we proposed to the use of the traditional search engine. The added value of lightweight ontologies instead of a hierarchy of concepts (as for example in the cat-a-cone system (Hearst 1997)) was estimated by asking astronomers to explore the theme ontology via two ways. First the interface was set to display only taxonomic relations and secondly all the semantic relations were shown. The instructions given to astronomers were to browse the theme ontology, stop on a concept they were not familiar with and read the articles that were associated to this concept. Both astronomers concluded that the context of concepts was much clearer when all the relations were stated. They could then evaluate more accurately what would be treated in the content of documents.

7 Conclusion

The model we have presented focuses on two aspects of the context: the theme of the search and the task for which the search is being made. We have chosen a model using lightweight ontologies. Based on this formalization, a sophisticated semantic indexing is implemented. The quality of the indexing process is also reinforced by our choice of using domain ontologies that provide more specific and precise knowledge representations than generic ones such as WordNet. This approach is interesting when theme domain ontologies exist. Numerous works have led to the development of domain ontologies in different fields such as medicine or in large enterprises (aerospace, car

manufacturing...). If such ontologies do not exist, we have developed an approach to build ontologies from texts and non-formal lexical resources but this is not part of this paper.

Moreover, contrary to existing approaches that consider the specific data interesting the user in his search, we have chosen to distinguish the modeling of the task and the modeling of the corpus theme. This distinction enables the reuse of our model for different tasks and different corpuses. The integration of the model in IRS is related to the semantic web, by a semantic indexing process that takes into account all the knowledge represented in the ontologies. Our model enriches traditional access to corpus content by giving a global view and specific views of the information available according to the two aspects of context. Evaluation has been made at different steps separately. Semantic indexing is found to be close to manual indexing. Regarding querying possibilities, we have considered different types of user information needs that cannot be handled by traditional IRS: the topics an author works on or the authors related to a given concept or set of concepts. Our system has been tested in the astronomy field for technical watch tasks. The system is currently being tested in the domain of e-learning, where the theme ontology refers to computer science and the task ontology to assimilating a lesson. The first results showed that students were particularly interested in having global and specific views on the lesson. They particularly appreciated visualizing the semantic links between the objects of their task (definition, exercise, computer science concepts dealt with). These results have emphasized the reusability of our approach. It can be used for any domain corpus and any search task involving specific meta-data. Our approach cannot be applied to multi-domain corpuses. We thus envisage integrating several theme ontologies in order to model such an environment that could be selected by the user according to his information need.

More traditional querying facilities (free text query) could be integrated easily. In fact, our semantic indexing already includes concept weighting that can play the role of usual term weighting. Query terms could be disambiguated thanks to the ontologies representing the context in order to determine the concepts referenced by the user, and matching documents could then be retrieved in a more traditional way (list of documents).

The model could be extended in several ways. Intelligent agents could be integrated on top of the model in order to assist the user with more sophisticated processes in his task. For example, analysis mechanisms such as those presented in (Dkaki 1997) could be integrated to the scientific watch task. These mechanisms allow for example to automatically extract co-authors by analyzing a set of representative documents. This process could be integrated in our system and presented to the user by an additional link between several concepts *Researcher*. Selecting one instance of researcher would then provide the correlated researchers.

We also plan to define and integrate user profiles. The profiles will allow the personalization of the browsing according to characteristics known about the user: presentation preferences, known theme concepts or predefined sub-ontologies, granules already explored.

8 Acknowledgments

The authors wish to thank Pascal Dubois and Andrea Preite Martinez, CDS astronomers, who carried out the evaluations presented in this article, as well as Didier Bourigault who provided Syntex and Upery system and Nathalie Aussenac for her valuable comments. Research outlined in this paper is part of the project WS-Talk that is supported by the European Commission under the Sixth Framework Programme (COOP-006026) and of the project "Astronomical Data Mass in Astronomy" founded by the Ministry for Research and Technology, within the framework of the call to projects 2003. However, views expressed herein are ours and do not necessarily correspond to the corresponding consortiums.

9 References

Allan J. (2003), Challenges in information retrieval and language modeling, SIGIR Forum, 37(1), pp 31-47.

Amous I., Chrisment C., Sèdes F. (2001) Extending the OOHDM methodology by eliciting metadata and generic structure, In Proceedings of DATAKON, pp 249-256.

Angelova G., Kalaydjiev O., Strupchanska A. (2004) Domain Ontology as a Resource Providing Adaptivity in eLearning, In Proceedings of the On the Move to Meaningful Internet Systems Workshop, LNCS 3292, pp 700-712.

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

Baeza-Yates R., Ribeiro-Neto B. (1999) Modern Information Retrieval, ACM Press, New York (NY).

Baziz M., Boughanem M., Aussenac-Gilles N., Chrisment C. (2005) Semantic Cores for Representing Documents in IR, In Proceedings of SAC, pp 1020-1026.

Benjamins R., Fensel D., Decker D., Gomez Perez A. (1999) (KA)2 : building ontologies for the internet : a mid-term report, In Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure, pp 1-24.

Bibow B, Szulman S, (1999) Terminae : a linguistics-based tool for building of a domain ontology, In D. Fensel and R. Studer, eds, Knowledge Acquisition, Modeling and Management, Proc. of the 11th European Workshop (EKAW'99), LNAI 1621, Springer-Verlag.

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

Cimiano P., Völker J. (2005) Text2Onto, A framework for Ontology Learning and data-driven Change Discovery, Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, Lecture Notes in Computer Science, n°3513 pp 227-238.

Croft, W. B., Thompson, R. H. (1987), I3R: a new approach to the design of document retrieval systems, In Journal of the American Society for Information Science , vol 38, pp 389--404

Desmontils E., Jaquin C. (2002) Indexing a Web site with a terminology oriented ontology, In Cruz S. Decker, J. Euzenat, D.L. McGuinness Eds, ISBN 1-58603-255-0, pp 181-197.

Dkaki T., Dousset B., Mothe J. (1997) Mining information in order to extract hidden and strategical information, In Proceedings of the International Conference on Computer Assisted Information Retrieval, pp 32-51.

Englmeier K., Mothe J., Murtagh F. (2004) Adapting the communication capacity of Web services to the language of their user community, In proceedings of the IEEE International Conference on Web Services, http://www.irit.fr/recherches/IRI/SIG/personnes/mothe/pub/ICWS04.pdf

Fernandez M., Gómez-Pérez A., Juristo N. (1997) METHONTOLOGY: from ontological art towards ontological engineering, In Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97) pp 33-40.

Freund L., Toms E.G. (2005) Contextual search: from information behaviour to information retrieval, In Proceedings of the Annual Conference of the Canadian Association for Information Science. http://www.cais-acsi.ca/proceedings/2005/freund_2005.pdf

Gangemi A., Guarino N., Masolo C., Oltramari A., Schneide L. (2002) Sweetening Ontologies with DOLCE, Source Lecture Notes In Computer Science; Vol. 2473 archive, In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, pp 166-181.

GuarinoN., Welty C (2002) Evaluating Ontological Decisions with OntoClean, In Communication of the ACM, 45(2), pp 61-65.

Guha R.V., McCool R., Miller E. (2003) Semantic search, In Proceedings of the 12th International World Wide Web Conference, pp 700-709.

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, pp 29-41.

Hernandez N., Mothe J. (2004) An approach to evaluate existing ontologies for indexing a document corpus, In Proceedings of the International Conference on Artificial Intelligence: Methodology, Systems, Applications - Semantic Web Challenges -, pp 11-21.

Mothe J., Hernandez N. (2006) TtoO: Mining thesaurus and texts to build and update a domain ontology, In Data Mining with Ontologies: Implementations, Findings, and Frameworks. H. O. Nigro, S. G. Císaro, and D.Xodo. Idea Group Inc.

Jiang J.J., Conrath D.W. (1997) Semantic similarity based on corpus statistics and lexical terminology, In Proceedings of the International Conference on Computational Linguistics, (RoclingX).

Jones G.J.F. (2004), The role of context in information retrieval, In proceedings of ACM SIGIR Workshop on Information Retrieval in Context, pp 20-23.

Kiryakov A., Popov B., Terziev I., Manov D., Ognyanoff D. (2004) Semantic annotation, indexing, and retrieval, Journal of Web Semantics, 2(1). Pp 49-79

Lassila O., McGuiness D. (2001) The role of frame-based representation on the semantic Web, Technical report, KSL-01-02, Knowledge Systems Laboratory, Stanford University.

Lord P.W., Stevens R.D., Brass A., Goble C.A. (2003) Semantic similarity measures as tools for exploring the Gene Ontology, In Proceedings of the Pacific Symposium on Biocomputing, pp 601-612.

McGuinness D.L, van Harmelen F. (2004) OWL Web Ontology Language Overview, W3C Recommendation http://www.w3.org/TR/owl-features/, 10 February 2004.

Maedche A., Staab S. (2002) Measuring similarity between ontologies, In Proceedings of the 13th International Conference EKAW, pp 251-263.

Mihalcea R., Moldovan D.I. (2000) Semantic Indexing using WordNet Senses, In Proceedings of ACL Workshop on IR & NLP.

Miller G.A. (1988) Nouns in WordNet, In WordNet, An Electronic Lexical Database, C. Fellbaum (Ed), pp 23-46, MIT Press.

Mothe J., Egret D., Chrisment C., Englmeier K. (2002) Knowledge discovery in bibliographic collections using concept hierarchies and visualisation tools, Library and Information Services in Astronomy, LISA IV, pp 233-241.

Resnik P. (1995) Using information content to evaluate similarity in a taxonomy, In Proceedings of the joint conference in Artificial Intelligence.

Robertson S.E., Sparck Jones K. (1976) Relevance weighting of search terms, Journal of the American Society for Information Sciences, 27 (3), pp 129-146.

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,4 (4).

Spärck Jones K. (1999) IR lessons for AI, In Proceedings of Searching for Information, Artificial Intelligence and Information Retrieval Approaches, IEEE Special event.

Staab S., Studer R. (2004) Handbook on Ontologies, Springer, International Handbooks on Information Systems, ISBN 3-540-40834-7.

Stuckenschmidt H., van Harmelen F., de Waard A., Scerri T., Bhogal R., van Buel J, Crowlesmith I., Fluit C., Kampman A., Broekstra J., van Mulligen E. (2004) Exploring large document repositories with RDF technology: the DOPE project, Intelligent system IEEE, 19(3), pp 34- 40.

Studer R., Benjamins R., Fensel D. (1998), Knowledge Engineering: Principles and Methods, Data and Knowledge Engineering, 25(1-2), pp 161-197.

Vallet D., M. Fernández, P. Castells (2005) An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, pp 455-470.

Taylor R.S. (1968) Question negotiation and information seeking in libraries. College and Research Libraries 29, pp 178-194.
Tudhope D., Alani H., Jones C. (2001) Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, Vol 41, pp1-8.
Turtle H.R. (1991) Inference Networks for Document Retrieval, PhD Thesis, University of Massachusetts.