D’un thesaurus vers une ontologie de domaine pour l’exploration d’un corpus

Claude Chrisment (*), Françoise Genova (**), Nathalie Hernandez(*), Josiane Mothe (*,*)**
chrisment@irit.fr, genova@cluster.u-strasbg.fr, hernandez@irit.fr, mothe@irit.fr

(*) IRIT, 118 route de Narbonne, 31040 Toulouse cedex , FRANCE,
(**) CDS, 11, rue de l'Université, 67000 Strasbourg, FRANCE
(***)IUFM, 56 av. de l’URSS, 31078 Toulouse cedex, FRANCE

Mots clefs :
Thesaurus, Ontologie, Veille, Exploration de corpus, méthodologie
Keywords:
Thesaurus, ontology, technological watch, science monitoring, exploration of a document collection, methodology
Palabras clave :
Escudriñar científico y tecnológico, ontologoas, exploración de una colección del documento, metodología.

Résumé

Dans cet article, nous proposons une méthodologie et des techniques associées pour permettre de transformer un thésaurus en une ontologie de domaine. L’originalité de notre approche repose sur le fait que l’ontologie est construite à partir de deux sources de connaissances non formalisées : celle issue du thésaurus et celle issue de documents du domaine. Nous avons appliqué notre démarche au cas de la transformation du thésaurus IAU de l’astronomie et les évaluations par des astronomes ont montré l’intérêt de nos propositions. L’utilisation de ressources de connaissances formalisées est un atout pour les systèmes qui doivent explorer des collections de documents de gros volumes.

1 Introduction

De nombreux thésaurus ont été créés dans différents domaines dans l'objectif de proposer un vocabulaire contrôlé pour l'indexation de ressources documentaires et pour l'aide à la formulation d'une requête par un documentaliste. Ils ont nécessité de lourds efforts pour leur conception manuelle. L'existence de normes (ISO 2788 et ANSI Z39) permet d'uniformiser leur contenu en termes de liens sémantiques entre unités lexicales (synonymie, liens hiérarchiques et d'association). Cependant, leur format n'est pas normalisé : fichiers ascii, html, bases de données co-existent. Pour faire face à ce problème, les normes en cours d'élaboration dans le cadre du W3C comme SKOS Core[1] visent à faire migrer les thésaurus vers des ressources disponibles sur le Web sémantique en se basant sur le langage OWL. La disponibilité de telles ressources sous format normalisé est un enjeu important dans le domaine de la veille et de l'accès à l'information [6]. La normalisation présente principalement trois avantages. Le premier est que l'uniformisation de leur représentation à partir de langages dédiés au web sémantique (tel que RDF et OWL) permettra à ces ressources d'être distribuées sur le web. De plus, ces ressources pourront être uniformément manipulées à partir d'outils dédiés aux ontologies pour leur visualisation, l'annotation, etc… Enfin, les processus de veille et d'accès à l'information pourront s'appuyer sur ces ressources élémentaires simples, sans avoir à faire face à l'hétérogénéité des formats. D'un point de vue de la représentation des connaissances, les thésaurus ont un faible degré de formalisation. Ce sont des collections de termes qui sont organisées suivant une ou plusieurs hiérarchies avec des relations entre termes. Les thésaurus n'ont pas de niveau d'abstraction conceptuelle [13]. La distinction entre un concept et sa lexicalisation n'est pas clairement établie. Les relations de synonymies sont établies entre les termes mais les concepts ne sont pas identifiés. Ceci s'explique par l'utilisation initiale des thésaurus, qui n'ont pas pour objectif de refléter comment le monde peut être compris en termes de sens mais en termes de terminologie et de catégories servant à l'indexation manuelle de documents d'un domaine. Pour réduire la complexité de leur élaboration, les concepteurs de thésaurus n'ont pas intégré ce niveau d'abstraction. De plus, la couverture sémantique des thesaurus est limitée. En effet, les relations entre termes sont vagues et ambiguës. Les liens sémantiques qu'ils contiennent reflètent parfois l'utilisation prévue du thésaurus plutôt que les liens sémantiques réels entre termes. Ces relations peuvent ainsi englober les relations « est une instance de » ou « est une partie de » [4]. La relation associative « est lié à » est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques [14]. Par exemple, dans le thésaurus BIT[2] relatif au monde du travail, le terme « famille » est lié aux termes « femme » et « congé familial », la relation sémantique entre ces deux paires de termes est intuitivement différente. Par les choix faits lors de leur conception, les thésaurus manquent de formalisation et de cohérence par rapport aux ontologies légères.

Les ontologies légères ou formelles ne posent pas ce type de problème. Elles sont supposées respecter la relation de subsomption dans l'organisation hiérarchique des concepts. D'autre part, les liens d'associations entre concepts sont sémantiquement mieux décrits. Cependant, leur élaboration est coûteuse ; elle nécessite de nombreuses interventions manuelles. En effet, les techniques de construction d'ontologies de la littérature basent généralement l'élaboration de l'ontologie sur aucune connaissance préalable du domaine.

Notre approche vise au contraire à réutiliser les thésaurus de domaine qui ont nécessité de lourds efforts de conception pour l'élaboration de nouvelles ressources d'un niveau formel plus élevé. La conception d'ontologies à partir de thésaurus présente l'avantage de reposer sur l'ensemble des termes qu'il contient et qui ont été identifiés par des experts comme étant représentatifs du domaine. Cependant, elle doit prendre en compte les différences fondamentales entre thésaurus et ontologie. La principale difficulté consiste à capturer la sémantique implicitement présente dans les thésaurus habituellement utilisés par des documentalistes.

En prenant en compte ces principales différences, nous proposons une méthode pour transformer un thésaurus en ontologie légère de domaine pour l'indexation de corpus en plusieurs étapes. Cette

[1] http://www.w3.org/TR/swbp-skos-core-guide/

[2] http://www.ilo.org/public/libdoc/ILO-Thesaurus/french/tr1740.htm

méthode vise à s'appliquer à n'importe quel thésaurus de domaine conçu sous les normes ISO 2788 et ANSI Z39. Ces thésaurus sont monolingues et ne sont pas organisés suivant des facettes.
Nous illustrons nos propositions à partir du thésaurus de l'astronomie IAU[3] ; les validations s'appuient également sur ce thésaurus.
La section 2 présente la méthode que nous proposons. Cette section explicite les problématiques auxquelles la méthode doit répondre, les différentes étapes qu'elle met en place, ainsi que le schéma conceptuel de l'ontologie choisi. Les sections suivantes décrivent les étapes de la transformation de l'ontologie. La section 4 présente les mécanismes utilisés pour créer le niveau d'abstraction conceptuel à partir du thésaurus. La section 5 explique comment la structure de l'ontologie est construite (liens entre concepts).

2 Présentation de la méthode

La méthode que nous proposons vise à permettre l'élaboration d'une ontologie légère de domaine pour l'exploration de corpus, à partir d'un thésaurus. Afin de capturer la sémantique implicitement présente dans le thésaurus et de mettre à jour la connaissance représentée à partir de la connaissance actuelle d'un domaine, la méthode se base sur l'analyse de documents textuels. Les problématiques auxquelles doit répondre la méthode sont situées dans le cadre général de la construction d'ontologies à partir de textes et repose sur différentes étapes.

2.1 Cadre général

La méthode que nous proposons s'appuie sur des documents textuels. La méthodologie TERMINAE [1] décrit les différentes étapes dans la construction d'une ontologie à partir de textes. Nous nous basons sur ces étapes pour spécifier la méthode permettant la transformation d'un thésaurus en ontologie. Afin d'identifier les éléments clés dans la transformation d'un thésaurus, nous reprenons les étapes de la méthodologie et les choix que nous faisons pour chacune d'entre elles.

- La première étape de la méthodologie TERMINAE vise à spécifier les besoins auxquels doit répondre l'ontologie. Dans le cas de la transformation d'un thésaurus en ontologie légère de domaine pour l'exploration de corpus, les besoins que nous identifions sont les suivants :
  - la spécification des termes du domaine et de leurs variantes lexicales afin de les détecter dans les granules documentaires,
  - le regroupement de ces termes en concepts afin de déterminer les objets et notions référencés dans les documents,
  - la structuration des concepts à partir de relations taxonomiques et associatives afin de permettre une indexation sémantique de qualité,
  - la formalisation de l'ontologie dans un langage interprétable par le système afin qu'il soit capable de la manipuler.

  Une ontologie légère créée pour l'exploration de corpus doit donc intégrer ces différents éléments.

- La deuxième étape repose sur le choix du corpus de référence à partir duquel l'ontologie est construite. Ce choix est un paramètre déterminant de l'élaboration de l'ontologie [3]. Le corpus doit décrire les éléments de connaissance qui seront intégrés dans l'ontologie. Dans le cas de la transformation d'un thésaurus, le corpus doit répondre à deux conditions. Il doit tout d'abord permettre de capturer la connaissance implicite qui n'est pas formalisée dans le thésaurus. Ensuite, le corpus doit aider à la mise à jour de la connaissance à partir de documents récents du domaine. L'ontologie étant créée pour des activités d'exploration de corpus, le corpus considéré doit aider à préciser le contexte associé à des documents du domaine d'intérêt considéré. Dans notre approche, le corpus est extrait de corpus existants et des experts doivent valider qu'il couvre l'ensemble du domaine sur une période représentative. Des résumés d'articles publiés dans des revues du

[3] http://www.site.uottawa.ca:4321/astronomy/index.html

domaine permettent de décrire ce type d'information. Les articles complets pourraient être utilisés mais l'avantage des résumés est que les informations qu'ils détiennent sont synthétisées.

- La troisième étape est celle de l'étude linguistique du corpus. Cette étape vise à extraire des documents les termes représentatifs du domaine et leurs relations (lexicales et syntaxiques) en utilisant des outils dédiés. A la fin de cette étape, on obtient un ensemble de termes, de relations entre ces termes et des regroupements. Dans le cadre de la transformation d'un thésaurus, cette étape intègre la connaissance représentée dans le thésaurus. Les termes présents dans le thésaurus sont représentatifs du domaine. Ils peuvent être regroupés à partir des relations du thésaurus. L'étude linguistique du corpus de référence est également nécessaire pour extraire les termes du domaine non présents dans le thésaurus et les relations entre termes qui n'y sont pas explicitées. Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique SYNTEX [2]. Cet analyseur à l'avantage de se baser sur un apprentissage endogène pour effectuer des analyses sur des corpus de différents domaines. Il permet d'extraire les syntagmes des documents ainsi que leur contexte d'apparition (mots qu'ils régissent et par qui ils sont régis). Une méthode additionnelle doit cependant être élaborée pour définir les mécanismes permettant de sélectionner les termes et leurs relations, à partir de la connaissance extraite du thésaurus et des informations extraites du corpus. La méthode que nous proposons vise à répondre à cette problématique.

- La quatrième étape correspond à la normalisation des résultats obtenus à l'étape précédente. A partir des termes et des relations lexicales, des concepts et des relations sémantiques sont définis. Au niveau de cette étape, le thésaurus peut être utilisé pour aider à la spécification des concepts.

- La dernière étape est celle de la formalisation, le réseau sémantique défini à l'étape précédente est traduit dans un langage formel. La formalisation de l'ontologie créée à partir d'un thésaurus peut être réalisée à partir du langage OWL [9]. Ce langage, au cœur du web sémantique a l'avantage d'être constitué de trois sous-langages d'un niveau de formalisation incrémentale. L'utilisation d'OWL-Lite permet une première formalisation de l'ontologie qui pourra évoluer. Ce langage permet de plus de représenter l'ensemble des éléments spécifiés par les besoins auxquels doit répondre une ontologie légère dans le cadre de l'exploration de corpus.

Pour la transformation d'un thésaurus, la méthode que nous proposons vise donc à mettre en œuvre les étapes 3, 4 et 5 spécifiés dans la méthodologie TERMINAE. Elle se base sur un mécanisme décomposé en différentes étapes décrites dans la section suivante.

2.2 Etapes de la méthode

La méthode proposée repose sur trois étapes. Ces étapes sont décrites dans la figure 1.

La première étape vise à extraire du thésaurus un ensemble de concepts ainsi que leurs variations lexicales. Peu de méthodes dans la littérature mettent en œuvre cette étape. La majorité d'entre elles considère en effet qu'un concept est référencé à partir d'un seul terme [7] [10] [11] [12]. Par l'utilisation d'un thésaurus, notre méthode vise à proposer un mécanisme automatique de regroupement des labels d'un même concept. Cette étape est décrite dans la section 3.

La deuxième étape permet de structurer les concepts de l'ontologie à partir de la détection de relations taxonomiques et associatives dans le thésaurus et dans le corpus. Cette étape soulève différentes problématiques de la construction d'ontologies. L'une d'elles relève de la difficulté à organiser les concepts par des relations taxonomiques. Dans notre cas, les relations hiérarchiques entre termes du thésaurus peuvent être utilisées pour aider à la détection de ces relations. Cependant, un des inconvénients des thésaurus est que le niveau hiérarchique le plus général est souvent composé de nombreux termes. Afin d'organiser les concepts à partir d'un niveau d'abstraction comportant un nombre limité de concepts, nous proposons l'utilisation d'une ontologie générique. Cette ontologie est utilisée pour définir semi-automatiquement les types abstraits du domaine et structurer l'ontologie. Le mécanisme développé est décrit dans la section 4. Une autre problématique que fait intervenir cette

étape est la détection de relations associatives entre concepts et la désignation de ces relations sémantiques. Peu de méthodes désignent correctement les labels des relations. Nous proposons un mécanisme visant à proposer semi-automatiquent ces relations ainsi que leur label. Le mécanisme repose sur l'analyse syntaxique du corpus de référence qui permet d'extraire les syntagmes constituant le lexique du corpus ainsi que le contexte dans lequel ils apparaissent (noms et verbes qu'ils régissent et par qui ils sont régis). Il est décrit dans la section 5.

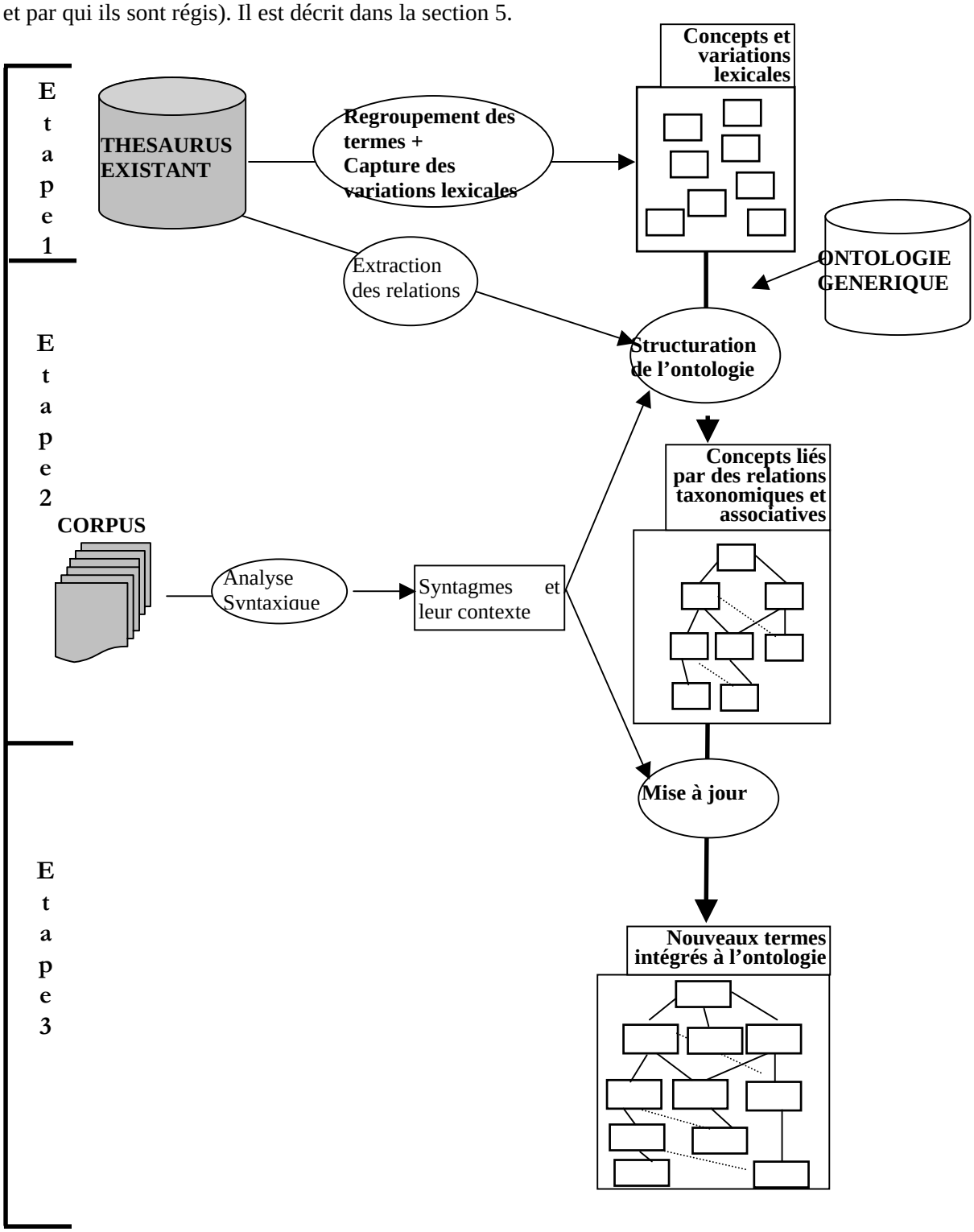

Figure 1 Etapes de la méthode

Contrairement à ce que préconise la méthodologie TERMINAE, la formalisation de l'ontologie est réalisée à la fin de ces différentes étapes après la validation des éléments proposés par un expert du domaine. Ce choix est justifié par le besoin, du concepteur de l'ontologie et de l'expert du domaine, de visualiser les éléments de connaissance jusque là représentés. Le schéma conceptuel utilisé et la formalisation qui lui est associée sont présentés dans la section suivante.

2.3 Schéma conceptuel

Le schéma conceptuel définit la structure de l'ontologie qui est élaborée. Cette structure doit faciliter la transformation d'un thésaurus traditionnel en une ontologie et permettre la représentation des éléments spécifiés par les besoins auxquels doit répondre l'ontologie. Il se veut simple pour permettre son adaptation à tous thésaurus respectant les normes ISO 2788 et ANSI Z39. Comme nous l'avons justifié plus haut, l'implantation de ce schéma repose sur des éléments spécifiés dans le langage OWL-Lite.

Le haut niveau conceptuel est présenté dans la figure 2 et est décrit dans les différents sous-paragraphes suivants.

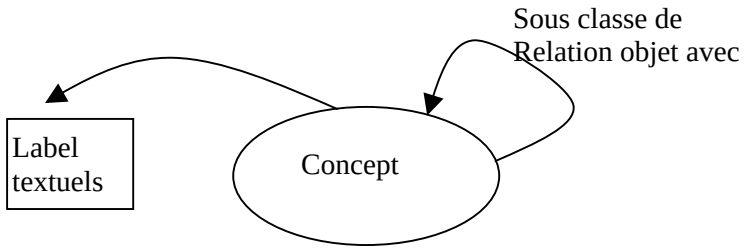

Figure 2 Haut niveau du schéma conceptuel de l'ontologie

2.3.1 Concept et label textuel

Un concept est représenté à partir d'une classe OWL `<owl:Class rdf:about="identifiant_unique">`. Elle est identifiée à partir d'un identifiant unique. Les labels d'une classe sont représentés par la propriété label `<rdfs:label>`. Dans le schéma conceptuel proposé dans [8], les labels sont de deux types : les labels représentant les termes principaux et ceux représentant les variations lexicales de ces termes. Cette approche peut être intéressante dans le cas où l'application et l'utilisateur doivent avoir une vision différente du contenu de l'ontologie. Dans la mesure où cette différenciation n'a pas lieu d'être dans la recherche de l'adéquation entre une ontologie et un corpus, ni dans le processus de RI, nous avons choisi de ne pas différencier les labels par rapport à leur rôle dans la désignation du concept. Les différentes variations lexicales des termes désignant le concept sont ainsi représentées par cette même propriété.

2.3.2 Relation entre concepts

Les concepts sont ensuite organisés à partir de relations taxonomiques représentées par la propriété `<rdfs:subClassOf>`.

Les concepts peuvent aussi être reliés entre eux à partir de relations non taxonomiques. Ce type de relations est représenté par l'intermédiaire de la propriété `<owl:ObjectProperty>` qui permet de lier deux concepts en spécifiant le concept de départ de la relation (`rdfs:domain`) et le concept d'arrivée (`rdfs:range`). Des propriétés peuvent être ajoutées à la relation, telles que la transitivité (`<rdf:type rdf:resource="&owl;TransitiveProperty"/>`), la symétrie (`<rdf:type rdf:resource="&owl;SymmetricProperty"/>`)et la fonctionnalité (`<rdf:type rdf:resource="&owl;FunctionalProperty"/>`), l'inverse d'une autre relation `<owl:inverseOf rdf:resource="#nom_propriété_inverse" />`.

2.3.3 Schéma conceptuel d'un thésaurus

L'ensemble des éléments du schéma conceptuel précédemment décrit ne sont pas présents dans un thésaurus. Un thésaurus est un ensemble de termes organisé suivant un nombre restreint de relations [5]. Les relations présentes dans un thésaurus répondant aux normes ANSI Z39 et ISO 2788 sont rappelées dans la figure 3.

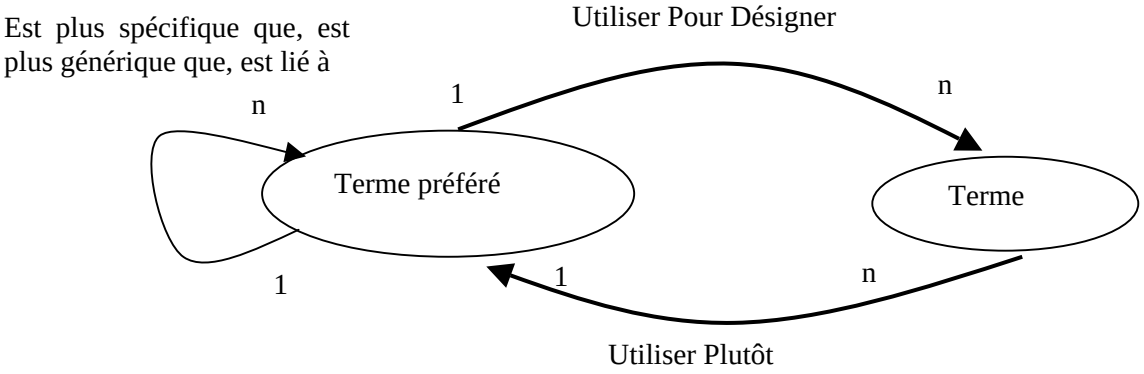

Figure 3 Rappel des relations entre termes dans un thésaurus

Dans notre méthode, nous considérons que les thésaurus réutilisés pour construire une ontologie sont de ce type.
Nous faisons trois hypothèses sur la réutilisation des relations entre termes :

- Les termes préférés sont les termes principaux du domaine et sont des indices pour constituer les termes désignant les concepts du domaine,
- Les relations entre termes et termes préférés sont des relations de synonymies entre termes, elles permettent de regrouper les termes comme étant label d'un même concept,
- Les relations entre termes préférés sont des indices pour définir des relations entre concepts.

A partir de ces hypothèses, nous définissons des méthodes permettant d'extraire les éléments du schéma conceptuel de l'ontologie d'un thésaurus et de documents textuels.

3 Conceptualisation du lexique du thésaurus

Cette étape vise à extraire du lexique du thésaurus une conceptualisation afin de formaliser un premier ensemble de concepts de l'ontologie.

3.1 Regroupement des termes en concepts

3.1.1 Regroupement basé sur les relations explicites UP et UPD

Afin d'extraire les concepts issus du lexique du thésaurus, les termes dits « préférés » ainsi que les relations du type « *Utiliser plutôt* » (UP) et *« Utiliser pour désigner » (UPD)* sont analysées. Nous interprétons ces relations comme des relations synonymies entre termes.

Des groupements de termes sont réalisés à partir de chacun des termes préférés et de l'ensemble des termes auxquels ils sont liés par les relations UP et UPD.

Si t3 UP t1 alors t1 et t3 sont regroupés, avec t1 terme préféré Si t1 UPD t2 alors t1 et t2 sont regroupés, avec t1 terme préféré **(R1)**

3.1.2 Regroupement basé sur la fermeture transitive des relations UP et UPD

Les groupements précédents sont ensuite agrégés à partir de la fermeture transitive des relations UP et UPD. Dans le cas où un terme préféré à l'origine d'un premier groupement apparaît dans un autre groupement, tous les termes liés au terme préféré et le terme préféré lui-même sont ajoutés aux groupements auxquels il est lié par une des relations.

La fermeture transitive consiste à regrouper les termes à partir de la règle R2.

Si t1 UPD t2 et t2 UPD t3, alors t1 UPD t3 => t1, t2 et t3 sont regroupés,
avec t1 terme préféré principal
Si t4 UP t5 et t5 UP t6 alors t4 UP t6 => t4, t5 et t6 sont regroupés,
avec t6 terme préféré principal
(R2)

La figure 4 schématise plusieurs exemples de groupements. Pour faciliter la lisibilité, les termes préférés sont en gras majuscules. Les termes regroupés par R1 sont soulignés en pointillé. Les termes regroupés par la règle R2 sont soulignés en trait plein.

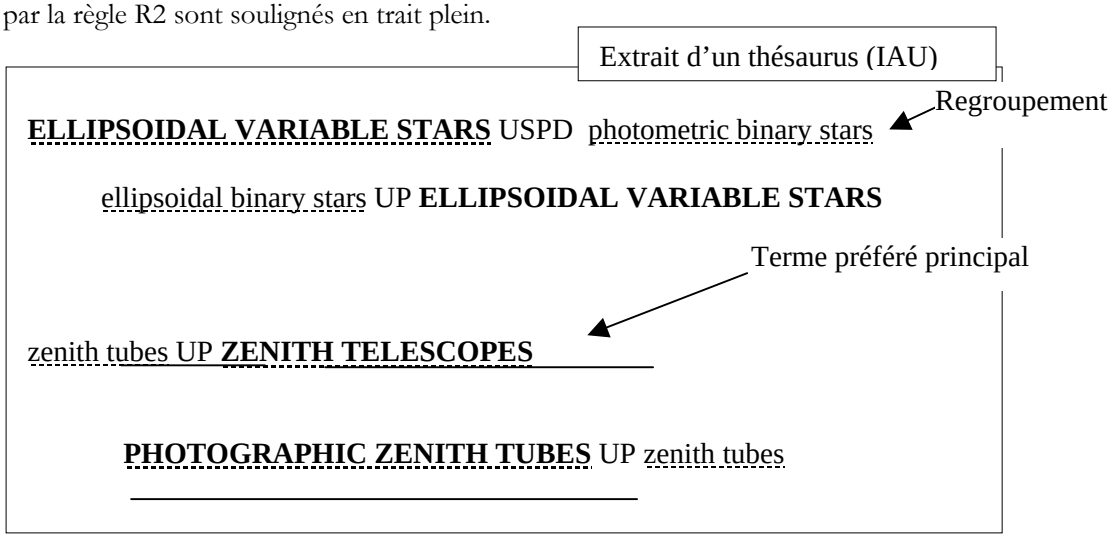

Figure 4 Exemples de groupements des termes du thésaurus

Les groupements de termes ainsi réalisés constituent l'ensemble des labels des futurs concepts de l'ontologie.

3.1.3 Identifiant du concept

L'identifiant d'un concept est déterminé par le terme préféré à l'origine du groupement. Le choix de ce terme comme identifiant permet de garder un lien entre la future ontologie et le thésaurus. Les identifiants des concepts correspondent ainsi à des entrées du thésaurus. Un terme peut être polysémique (label de plusieurs concepts) dans le cas où il était lié dans le thésaurus à deux termes préférés distincts.

Si t1, t2,… et tn regroupés avec t1 terme préféré principal
=> création du concept c d'identifiant t1 et de labels t1, t2,… et tn
(R3)

3.2 Capture des variations lexicales

La forme lexicale sous laquelle se trouvent les termes du thésaurus est un sujet délicat et largement détaillé dans l'ensemble des normes dédiées au thésaurus [ISO 2788, AFNOR NF Z47-100, ANSI Z39]. Ceci s'explique par l'ambiguïté posée par le rôle des termes dans un thésaurus. Les termes peuvent soit représenter des catégories d'objets similaires, soit désigner le sens des objets. Dans le cas où le terme représente une catégorie, le pluriel du terme est préféré et, dans le cas où le terme définit le sens du terme, le singulier est choisi. Les normes ISO 2788 et ANSI Z39 proposent, pour différencier ces cas de figure, la distinction des termes à partir de leur type : les termes désignant des objets dénombrables et les termes désignant des objets indénombrables. Lorsque peut être posée la question « combien d'objets représentés par le terme existent ?», le terme est intégré dans le thésaurus au pluriel, dans le cas contraire il l'est au singulier. Ces règles sont scrupuleusement respectées dans la plupart des thésaurus, comme dans le thésaurus de l'astronomie IAU. Il est cependant possible de trouver des variantes de l'application de ces règles. Par exemple, dans les thésaurus BIT http://www.ilo.org/public/french/support/lib/indexati/unit1/unit1.htm (Terminologie du travail, de l'emploi et de la formation), British Museum http://www.mda.org.uk/bmobj/Objintro.htm et Alcohol and Other Drug Thesaurus http://etoh.niaaa.nih.gov/AODVol1/titlepage.htm, les termes sont au singulier sauf si l'usage impose le pluriel.

Dans les ontologies, les termes sont utilisés pour référencer des concepts et décrire le sens associé aux objets qu'ils représentent. Il est donc important que les labels de l'ontologie ne représentent pas des catégories mais des unités de sens. Les termes doivent donc être au singulier.

Des techniques de lemmatisation ou le recourt à un expert peuvent être utilisées. Alternativement, une ressource lexicale telle que WordNet peut être utilisée. La figure 5 illustre les concepts identifiés dans la figure 4 pour lesquels les labels sont mis au singulier grâce à WordNet.

CONCEPT
Identifiant : ELLIPSOIDAL VARIABLE STARS
Labels :
ellipsoidal variable star
photometric binary star
ellipsoidal binary star

CONCEPT
Identifiant : ZENITH TELESCOPES
Labels :
zenith telescope
zenith tube
photographic zenith tube

Figure 5 Exemples de concepts labellisés par des termes au singulier

4 Construction de la structure de l'ontologie

La structure de l'ontologie définit les relations entre concepts établis suite aux étapes présentées dans la section précédente. La structure comprend des relations taxonomiques de type « est un » et des relations associatives qui sont obtenues par les méthodes décrites ici.

4.1 Construction de la hiérarchie de concepts

Certains liens hiérarchiques entre concepts sont directement issus des liens explicites présents dans le thésaurus. Des niveaux hiérarchiques supérieurs y sont ajoutés à partir de l'analyse des têtes et expansions des labels des concepts et de la création de types abstraits. La figure 6 schématise ces différents mécanismes.

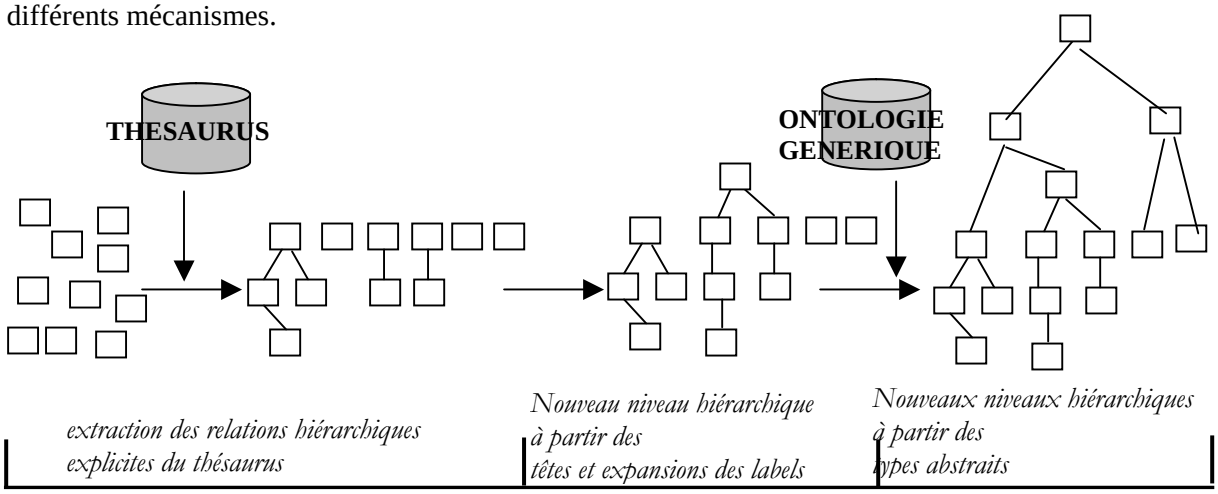

Figure 6 Mécanisme de construction de la hiérarchie de concepts

4.1.1 Extractions des relations hiérarchiques explicitées dans le thésaurus

Les concepts sont d'abord organisés hiérarchiquement à partir de la relation « sous classe de » du schéma conceptuel de l'ontologie. Afin d'extraire ce type de relation du thésaurus, les relations « est plus spécifique que » et « plus générique que » du thésaurus sont prises en compte. L'ensemble de ces relations définies pour les termes, devenus maintenant labels d'un concept, est retenu comme relations candidates pour représenter des relations « sous classes » entre le concept et le concept auquel se rapporte le terme lié dans le thésaurus. Les relations candidates doivent ensuite être analysées avec précaution car elles peuvent englober des relations de type « partie de » ou « instance de ». Nos travaux ne proposent pas de méthode automatique pour réaliser cette désambiguïsation. Il faut noter que beaucoup de thésaurus de domaine prennent la peine de considérer les relations « est plus spécifique que » et « plus générique que » de façon stricte. Cela est également le cas pour le thésaurus de l'astronomie IAU qui sert de validation à notre approche.

Si t1 est plus spécifique que t2 avec t1 label du concept c1 et t2 label du concept c2
=> c1 « est une sous classe de » c2
(R4)

4.1.2 Suppression de la redondance dans les relations hiérarchiques

Les thésaurus n'étant pas formalisés, des redondances dans la structure hiérarchique de l'ontologie construite avec les règles de R1 à R4 peuvent exister. La relation de généricité est une relation transitive, et permet le type d'inférence suivant : si A « est une sous classe de » B et B « est une sous classe de » C, alors A « est une sous classe de » C, A,B,C étant des concepts. La figure 7 en présente un exemple ; les flèches entre les rectangles représentant les concepts symbolisant la relation « est une

sous classe de ». Par la propriété de transitivité de la relation « est une sous classe de », la relation « est une sous classe de » entre *palenetary nebula* et *nebula* est donc inutile.

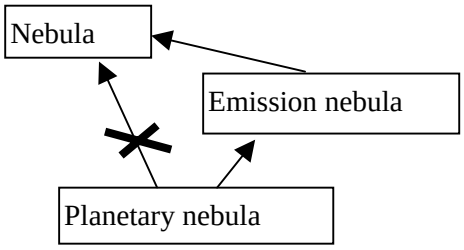

Figure 7. Exemple de redondance dans la hiérarchie de l'ontologie

Afin de supprimer les relations redondantes, la pertinence de chacune des relations «est une sous classe de » est vérifiée.
La suppression de la redondance est formalisée par la règle R5.

Pour tout concept $c \in C$,
Si $\forall c_i \in C\ c \neq c_i$, $\exists$ chem1, chem2 tel que chem1=chemin(c,c_i) et chem2=chemin(c,c_i), avec
chem1 $\neq$chem2
=> suppression de l'arc à l'origine du chemin le plus court
(R5)

4.1.3 Nouveaux niveaux hiérarchiques

Une des lacunes des thésaurus est que leur plus haut niveau hiérarchique contient généralement un très grand nombre de termes [13]. Ces termes sont ceux pour lesquels aucune relation « est plus spécifique » n'a été définie. Ceci s'explique par le fait que les thésaurus ne définissent pas de catégories génériques permettant de répertorier l'ensemble des termes du domaine. Cette même lacune est constatée dans les ontologies obtenues par la transformation d'un thésaurus. Ceci pose problème lorsqu'un utilisateur ou une application choisit d'explorer l'ontologie par une navigation de haut en bas. Le grand nombre de concepts du premier niveau rend le départ de sa navigation délicate. Par exemple, le niveau hiérarchique le plus générique de l'ontologie extraite du thésaurus IAU à cette étape de la transformation contient 1132 concepts.

Nous proposons donc l'ajout de niveaux hiérarchiques plus génériques qui facilitent la navigation dans l'ontologie. D'autre part, nous proposons la définition de concepts génériques (ou types abstraits) permettant de caractériser les concepts. Un concept générique ou abstrait fait référence à une notion abstraite et n'admet pas d'instance. Il est soit un véritable concept du domaine, soit un concept ajouté pour structurer la représentation. Dans [13], les concepts génériques sont définis à partir d'un schéma de catégorisation de haut niveau existant dans le domaine. Les concepts du plus haut niveau de l'ontologie sont liés manuellement aux concepts de ce schéma. Ce procédé ne peut pas être appliqué à tous les domaines, car de tels schémas n'existent pas toujours. De plus, il demande un travail manuel à l'expert qui doit affecter les milliers de classes de l'ontologie à l'une des centaines de classes du schéma. Nous proposons donc une autre approche plus automatisée.

4.1.3.1 Premier niveau de généralisation : tête et expansion des syntagmes

Pour créer un premier niveau d'abstraction, les concepts sont regroupés à partir de la tête des termes de leur label. Cette approche est suivie dans OntoLearn [15] pour créer la hiérarchie de concepts. Les concepts ayant des labels comportant la même tête sont définis comme étant des sous classes du concept labellisé par la tête (règle R6 et figure 8). Si ce concept n'existe pas dans l'ontologie, il est crée et appartient au nouveau niveau 0 de l'ontologie (règle R7 et figure 9). Ce mécanisme permet de créer un nouveau premier niveau de la hiérarchie contenant un nombre plus réduit de concepts.

Si tete(F^{-1} (c_1))= tete(F^{-1} (c_2)) alors si tete(F^{-1}(c_1))$\in$ L_{Onto}
=> c_1 « est une sous classe de » F(tete(F^{-1} (c_1)))
et c_2 « est une sous classe de » F(tete(F^{-1} (c_1)))

(R6)

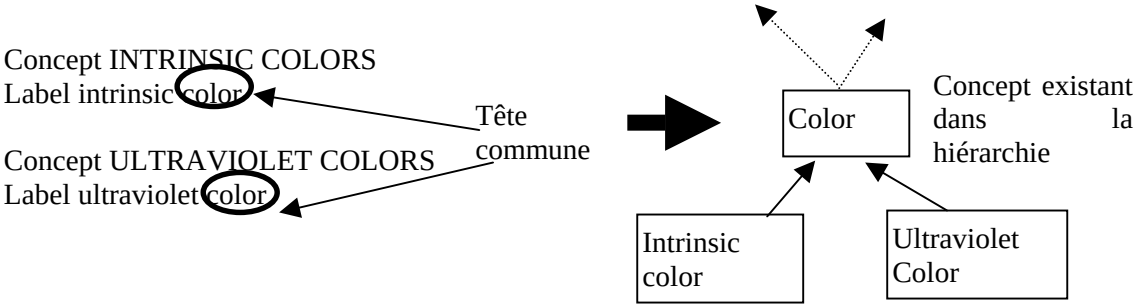

Figure 8. Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie

Si tete(F^{-1}(c_1))= tete(F^{-1} (c_2)) alors si tete(F^{-1} (c_1))$\in$ L_{Onto}
=> tete(F^{-1} (c_1)) est un nouveau concept c $\in$ C_{Onto} de label tete(F^{-1} (c_1)).
Il est ajouté à l'ontologie avec c_1 « est une sous classe de » c
et c_2 « est une sous classe de » c

(R7)

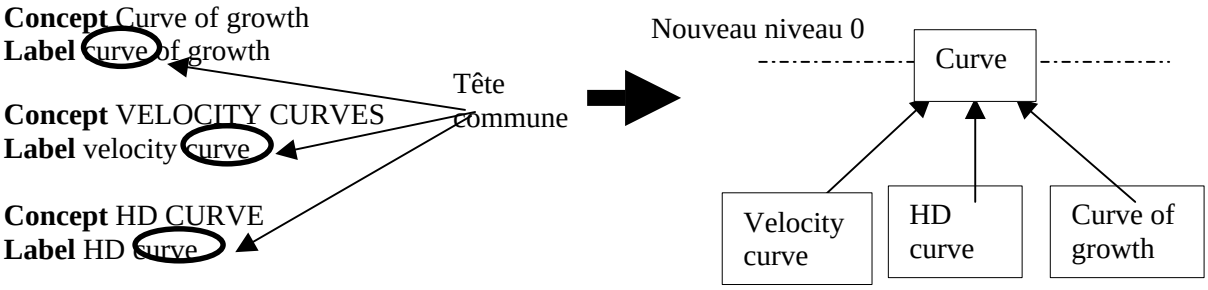

Figure 9. Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

4.1.3.2 Deuxième niveau de généralisation : types abstraits

La définition des types abstraits vise à identifier les concepts génériques dont dépendent les concepts du niveau 0 de généralisation précédent. Cette définition comporte deux étapes. Dans un premier temps, il s'agit de définir les types abstraits du domaine, puis de les associer aux concepts. La règle R8 synthétise les étapes qui sont décrites ci-dessous.

Si c $\Leftrightarrow$ sw avec sw$\in${synsets_WordNet}
=> c « est sous classe de » ta
Avec ta (type abstrait) est le plus spécifique hyperonyme de sw

(R8)

- *Définition des types abstraits*

Afin de définir ces types de façon automatisée, une ontologie de haut niveau comme par exemple WordNet ou DOLCE, est utilisée. Tout d'abord, les concepts de niveau 0 doivent être mis en correspondance avec les concepts de l'ontologie. Les types abstraits sont alors définis à partir des concepts les plus génériques associés aux concepts détectés. Nous décrivons dans cette section l'utilisation de WordNet pour expliciter cette étape.

Concernant la mise en correspondance des concepts de niveau 0 avec les synsets de WordNet, les labels des concepts de l'ontologie en cours de construction sont comparés aux entrées de WordNet. Chaque synset ainsi détecté est candidat pour représenter le concept dans WordNet. Dans le but de limiter les synsets extraits aux synsets se rapportant effectivement aux concepts de l'ontologie, un mécanisme de désambiguïsation est mis en place. Il prend en compte quatre éléments :

- le glossaire fourni par WordNet pour décrire en langage naturel le sens du synset
- les synsets descendants du synset en question par la relation hyperonymie dans Wordnet
- les synsets ancêtres du synset en question dans WordNet par la relation hyponymie dans WordNet
- les labels des concepts descendants du concept dans l'ontologie par la relation « est sous classe de »

Lorsque plusieurs synsets correspondent à un label d'un concept de niveau 0, le synset choisi est obtenu par trois méthodes de désambiguïsation qui sont mises en oeuvre séquentiellement :

(1) Les termes très généraux décrivant le domaine traité par l'ontologie sont tout d'abord spécifiés avec des experts du domaine. Ils sont ensuite recherchés dans le glossaire associé par WordNet à chacun des synsets candidats. Par exemple, le terme recherché dans le glossaire pourrait être « astronomie ». Si un de ces termes est retrouvé, le synset candidat est automatiquement choisi. Sinon, la méthode (2) est appliquée.

(2) Les synsets fils du synset sont comparés aux concepts fils du concept dans l'ontologie. Si au moins un des labels se rapportant aux concepts fils est retrouvé dans les synsets fils, alors le synset est choisi. Sinon, la méthode (3) est appliquée.

(3) Les synsets ancêtres du synset candidat sont analysés par la proposition (1). Un synset candidat est choisi dans le cas où la proposition est vérifiée, et, dans le cas contraire, le concept n'est pas associé à un synset de WordNet car aucun synset n'a pu être désambiguïsé.

Concernant l'identification des types, les synsets les plus génériques (i.e. les plus lontains ancêtres) des synsets désambiguïsés sont proposés pour représenter les concepts génériques de l'ontologie. Ils sont ensuite validés par un expert et intégrés à l'ontologie en tant que nouveaux concepts.

- *Association des concepts aux types abstraits*

Pour les concepts de niveau 0 de l'ontologie ayant été liés à un synset désambiguïsé, un lien est établi entre le concept et le type abstrait correspondant. Le lien est représenté dans l'ontologie en définissant le concept comme sous classe du type abstrait.

Dans le cas où la désambiguïsation n'a pu avoir lieu ou que les labels du concept n'étaient pas dans WordNet, l'association concept/type abstrait est réalisée manuellement.

La figure 10 présente des exemples de types abstraits extraits pour notre cas d'application.

Property : a basic or essential attribute shared by all members of a class
Phenomenon : any state or process known through the senses rather than by intuition or reasoning
Event : *something that happens at a given time*
Science : a particular branch of scientific knowledge
Instrumentation : an artifact (or system of artifacts) that is instrumental in accomplishing some end
Substance : that which has mass and occupies space
Relation : an abstraction belonging to or characteristic of two entities or parts together
Location : a point or extent in space

Angle : the space between two lines or planes that intersect; the inclination of one line to another
Plane : an unbounded two-dimensional shape
Region : the extended spatial location of something;
Object : a tangible and visible entity
Natural object : an object occurring naturally; not made by man
Artefact : a man-made object taken as a whole

Figure 10. Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

5 Détection des relations associatives

La deuxième étape dans la formalisation de la structure de l'ontologie vise à définir des relations associatives entre concepts de l'ontologie. Ces relations sont tout d'abord extraites des relations du thésaurus. De nouvelles relations entre les concepts sont ensuite extraites à partir de l'analyse du corpus de référence. Nous présentons dans cette section ces différents éléments.

5.1 Spécification de relations entre types abstraits

La spécification des relations sémantiques entre types abstraits de l'ontologie est fondée sur la proposition de relations associées à chaque type par une analyse syntaxique automatique du corpus de référence. Ces propositions servent de base à la définition manuelle de relations entre paires de type abstrait et sont synthétisées dans la règle R9.

Soient ta_1 et ta_2 deux types abstraits avec $ta_1 \in C_{Onto}$ et $ta_2 \in C_{Onto}$
Soient $r,r' \in R_{Onto}$ avec σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine
Si $r'(c_1,c_2)$ avec $c_1 \in C_{Onto}$ et $c_2 \in C_{Onto}$ et c_1 « est sous classe de » ta_1 et c_2 « est sous classe de » ta_2 et $G^{-1}(r')$=« est lié à »
$\Rightarrow G^{-1}(r') \in G^{-1}(r)$

(R9)

5.1.1 Proposition de relations

A partir de l'analyse syntaxique réalisée sur le corpus de référence, le contexte des labels de chacun des concepts est extrait. Nous entendons par contexte, les syntagmes dont les labels sont tête ou expansion, les compléments d'objet et les sujets de verbes dans lesquels les labels apparaissent. Ces contextes sont ensuite regroupés à partir des types abstraits auxquels se rapportent les concepts. Les termes apparaissant fréquemment dans les contextes regroupés sont retenus pour caractériser le type abstrait et servir de proposition aux labels des relations associatives que ses concepts fils peuvent avoir. Prenons, pour illustrer cette idée, le cas des contextes des concepts dépendant du type abstrait *instrumentation* dans l'ontologie de l'astronomie. Les termes apparaissant le plus fréquemment sont les verbes anglais « observe » et « measure ». Ces termes indiquent que les instruments astronomiques sont utilisés pour observer ou mesurer les autres concepts du domaine.

5.1.2 Définition de relations entre types

La définition des relations sémantiques est réalisée entre chaque paire de type abstrait. Une matrice à double entrée est ensuite réalisée. Cette matrice contient en ligne et en colonne l'ensemble des différents types abstraits identifiés manuellement sur la base des propositions précédentes. Chaque case de la matrice contient les relations possibles. Un extrait de la matrice proposée pour le domaine de l'astronomie est présenté dans le tableau 1. Il est important de noter que la diagonale de la matrice témoigne de relations particulières. Elles relient en effet des concepts de même type. Une proposition particulière est donc ajoutée pour ce type de relation, la proposition est la relation « partie de ». Les concepts étant de même type, ils peuvent avoir été liés parce que l'un d'eux spécifie une partie de l'autre. Sur la base des propositions précédemment faites, un expert du domaine identifie les relations

qui peuvent lier les concepts génériques deux à deux et reporte les labels qu'il choisit dans les cases de la matrice.

	Property	**Phenomenon**	**Event**	**Science**	**Natural object**	**Instrumentation**
Property	*Influences* *Is influenced by* *Determined by* *Determines* *Exclude* *Has part* *Is part*	*Is a property of* *induces*	*Is a property of* *induces*	*Is studied by*	*Is a property of*	*Is maked by* *Is observed by*
Instrumentation	Makes Observes	*Observes* *Measures*	*Observes* *Measures*	*Is Used to studied*	*Is observed by*	*Is ou has part* *exclude*

Tableau 1 Extrait de la matrice des relations entre types abstraits

5.1.3 Association des relations vagues du thésaurus et des relations entre type

Les relations vagues du thésaurus « est lié à » sont d'abord retranscrites dans l'ontologie. Ainsi, deux termes liés dans le thésaurus donneront lieu à une association entre les concepts dont ils sont labels dans l'ontologie. Cette association est ensuite spécifiée grâce aux relations identifiées dans la matrice entre les types abstraits associés à ces concepts. Par exemple, la relation identifiée entre les types abstraits « instrumentation » et « natural object» étant la relation « *observes* », la relation « est lié à » du thésaurus entre « *coronagraph* » et « *solar corona* » (concepts issus de ces deux types) est modifiée en la relation « *coronagraph* » « *observes* » « *solar corona* ». Si plusieurs relations sémantiques sont identifiées, le choix est laissé à l'expert du domaine.
Le mécanisme mis en place peut s'apparenter à celui proposé dans [Sorgel 2004]. Les relations entre concepts sont en effet établies à partir de l'analyse des relations du thésaurus et de la définition de patrons permettant de retrouver les relations sémantiques spécifiées dans l'ensemble du corpus. Plutôt que d'avoir à spécifier individuellement les relations vagues dans le thésaurus entre termes, l'expert doit seulement valider ou invalider les propositions qui lui sont faites sur la base de l'analyse du corpus et des relations entre les types abstraits. Ainsi, l'analyse que nous mettons en place facilite le travail de l'expert.

5.2 Détection de nouvelles relations associatives

Contrairement aux approches de la littérature visant uniquement à transformer un thésaurus en ontologie à partir de la connaissance représentée dans celui-ci, nous proposons d'établir de nouvelles relations associatives entre les concepts à partir de l'analyse de documents textuels du domaine (cf règle R10).

Sur la base de la matrice précédemment établie, de nouvelles relations sont décelées entre les concepts de l'ontologie. Pour cela, le contexte des différents labels des concepts dans le corpus est analysé. Deux approches sont utilisées pour considérer le contexte.

La première prend en compte les termes qui occurrent fréquemment autour des labels de concepts de l'ontologie.

La seconde se base sur l'analyse distributionnelle réalisée par le module UPERY de SYNTEX [2]. Ce type d'analyse consiste à rapprocher des syntagmes en fonction de la ressemblance de leur contexte. Les syntagmes déduits de l'analyse syntaxique sont rapprochés s'ils sont formés autour de la même relation et des mêmes têtes et queues. Par exemple, en considérant les syntagmes « star » « galaxy », « star mass » et « galaxy mass », les syntagmes « star » et « galaxy » sont rapprochées par le contexte « mass ». UPERY permet de rapprocher des syntagmes à partir d'un poids de proximité. Ce poids

prend en compte la productivité d'un terme et la productivité d'un concept. A partir d'un seuil fixé empiriquement sur ce poids, le module détecte des relations entre syntagmes mais ne désigne pas la relation sémantique qui les relie. Pour proposons donc d'utiliser les résultats de ce module pour la détection de nouvelles relations associatives qui sont typées par l'intermédiaire de la matrice.

Lorsqu'un label apparaît dans le contexte d'un concept ou les termes qui lui sont associés par l'analyse distributionnelle et qu'aucune relation ne lie les deux concepts dans l'ontologie, une relation est proposée entre les deux concepts. Cette relation prend en compte le type des deux concepts et est établie à partir de la matrice élaborée à l'étape précédente.

Par exemple, dans le contexte du label « *luminosity* » référençant le concept de même nom, le label « *galaxy* » correspondant au concept « *galaxy* » est retrouvé. Ces concepts étant de type « *property* » et « *natural object* », la relation « has a » est proposée entre « *galaxy* » et « *luminosity* » (cf tableau 1). Aucune relation n'ayant été précédemment établie entre ces deux concepts, la nouvelle relation est ajoutée à l'ontologie.

Soient ta_1 et ta_2 deux types abstraits avec $ta_1 \in C_{Onto}$ et $ta_2 \in C_{Onto}$
Soient r,r' $\in R_{Onto}$ avec σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine
Si $r'(c_1,c_2)$ décelée par l'analyse du corpus avec $c_1 \in C_{Onto}$ et $c_2 \in C_{Onto}$
=> $G^{-1}(r') \in G^{-1}(r)$

(R10)

6 Conclusion

Le procédé de transformation d'un thésaurus en ontologie légère que nous proposons repose sur trois étapes principales : l'extraction d'information du corpus, l'identification des concepts issus du thésaurus, la construction de structure de l'ontologie (hiérarchie de concepts et relations associatives entre concepts). Les procédés sont simples à mettre en œuvre et permettent d'extraire une ontologie légère. Ils nécessitent une validation par un expert du domaine, mais le travail qui lui est demandé est allégé par la proposition d'éléments à chacune des étapes. Le travail demandé à l'expert est moins important que celui demandé par les approches proposées dans [13] [16] car son travail consiste uniquement à valider les propositions. Contrairement aux approches présentées dans la littérature, le procédé mis en place vise non seulement à transformer le thésaurus mais aussi à intégrer de nouvelles connaissances dans l'ontologie (ajout de relations entre concepts).
Une contribution importante de notre travail est la proposition permettant de déceler puis de labelliser les relations associatives entre concepts. Elle repose sur la notion de type abstrait qui sont des concepts de haut niveau d'abstraction. La définition de relations sémantiques, validée par des experts est rapide, compte tendu du nombre limité de types abstraits. Ces relations permettent d'inférer des relations au niveau des concepts de plus bas niveau, en les associant à l'analyse syntaxique du corpus.
Cette méthodologie est bien adaptée lorsque le thésaurus initial est construit en respectant la sémantique de la relation « est un ». En revanche, et comme nous l'avons souligné précédemment, lorsque ce n'est pas le cas, une étape supplémentaire doit être ajoutée afin de distinguer les différentes relations telles que « est une partie de » ou « est une instance de ».
Les premières évaluations de nos propositions dans le cadre de l'astronomie a débuté. Les experts du domaine sont satisfaits des résultats. La pertinence des concepts créés et définis à partir de plusieurs labels a été validée par les astronomes. La validation a montré que la totalité des concepts créés étaient pertinents et que pour 85 % d'entre eux l'ensemble des labels était correct. Pour 15 % les labels ne sont pas corrects car ils se rapportent à des sous concepts des concepts pour lesquels ils sont définis. Ces labels ont donc été supprimés et ont mené à la création de nouveaux concepts définis comme sous concepts des concepts auxquels ils étaient rattachés à l'origine. L'organisation hiérarchique des concepts est réalisée par les règles R4 et R5. Les relations sont définies à partir de la relation « est plus spécifique » « est plus générique » du thésaurus. Ces relations ont mené à la définition de 2882 relations « sous classe de » dans l'ontologie. Parmi celles-ci, 193 relations redondantes ont été trouvées. Elles ont donc été supprimées de la hiérarchie de concepts. Les relations étant définies dans les spécifications du thésaurus pour ne comprendre que des relations du type « est plus spécifique »

« est plus générique », seules 5% de ces relations ont été analysées. La totalité d'entre elles a été validée.

Les perspectives de ce travail sont multiples. Concernant le contenu de l'ontologie, une première perspective concerne sa mise à jour. Les thésaurus reflètent des connaissances dans un domaine à un instant donné. Il est important que cette connaissance puisse être mise à jour. Nous travaillons donc sur la mise à jour d'une ontologie de domaine légère, en s'appuyant sur la connaissance qui peut être extraite des corpus (nouveaux concepts, nouvelles relations). Une autre perspective de ce travail concerne l'utilisation concrète d'une ontologie dans le domaine de la veille. En réalité, nous avons proposé un système d'exploration de corpus (OntoExplo) basé sur des ontologies de domaine [6]. Ce travail s'appuyait sur des ontologies pré-existantes. Dans cet article au contraire, nous nous sommes attachées à montrer comment une telle ressource pouvait être construite.

7 Remerciement

Les travaux présentés dans ce papier ont été réalisés dans le cadre des projets WS-Talk "WS-Talk: Web services communicating in the language of their user community" supporté par le Sixth Framework Programme of the European Community (2002-2006), COOP-CT-2004 006026 et le projet Masse de Données en Astronomie supporté par le ministère délégué à la Recherche et aux Nouvelles Technologies. Nous tenons à remercier particulièrement les astronomes du CDS qui ont évalué nos propositions.

8 Bibliographie

[1] N. Aussenac-Gilles, B. Biébow, S. Szulman, Modélisation du domaine par une méthode fondée sur l'analyse de corpus, dans les actes de la conférence IC'2000, Journées Francophones d'Ingénieurie des connaissances, pp 93-103, 2000.

[2] D. Bourigault, Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy, 2002, pp. 75-84

[3]. A. Condamines, Sémantique et Corpus, Hermès science publications, ISBN 2-7462-1055-X, 2005.

[4] D. H. Fischer, From Thesauri towards Ontologies?, in: el Hadi, Maniez & Pollitt (Eds.): Structures and Relations in Knowledge Organization, dans 5th Int. ISKO Conference, Lille, France, 1998. Würzburg: Ergon, pp. 18-30, 1998.

[5] D.J. Foskett, Thesaurus, In Encyclopedia of Library and Information Science, A. Kent, H. Lancour (Eds), p.416-463, 1980.

[6] N. Hernandez, J. Mothe, Ontologies pour l'aide à l'exploration d'une collection de documents, Veille Stratégique Scientifique & Technologique Systèmes d'information élaborée, Bibliométrie, Toulouse, Novembre 2004.

[7] A. Maedche and S. Staab. Mining ontologies from text. In Proceedings of EKAW-2000, Springer Lecture Notes in Artificial Intelligence (LNAI-1937), Juan-Les-Pins, France, 2000. Springer, 2000.

[8] A. Miles, D. Brichley, SKOS Core GuideW3C Working Draft 10 May 2005, http://www.w3.org/TR/swbp-skos-core-guide/

[9] D. L McGuinness, F. Van Harmelen, OWL Web Ontology Language Overview, W3C Recommendation http://www.w3.org/TR/owl-features/, 10 février 2004.

[10] E. Morin, C. Jacquemin, Projecting Corpus-Based Semantic Links on a Thesaurus, Dans 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, Juin 1999.

[11] Sang Ok Koo, Soo Yeon Lim, Sang Jo Lee Building an Ontology Based on Hub Words for Information Retrieval IEEE/WIC International Conference on Web Intelligence (WI'03) ,October 13 - 17, 2003 Halifax, Canada.

[12] M. Sanderson, W.B Croft., Deriving concept hierarchies from text, in Proceedings of the 22nd annual conference ACM SIGIR, pp 206-213, 1999.

[13] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer and S. Katz, Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information, Volume 4 Issue 4, Article N° 257, 2004

[14] D. Tudhope, H. Alani, C. Jones, Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, Volume 1 Issue 8, Article No. 41, 2001.

[15] P. Velardi, P. Fabriani, M. Missikoff: Using text processing techniques to automatically enrich a domain ontology, FOIS, pp 270-284, 2001:

[16] B. Wielinga, G. Schreiber, J. Wielemaker, and J. A. C. Sandberg. From thésaurus to ontology. Internation Conference on Knowledge Capture, Victoria, Canada, Octobre 2001.