Enrichissement sémantique pour la recherche d'information : méthodologie de transformation d'un thésaurus en une ontologie de domaine

Claude Chrisment *, Nathalie Hernandez *,†, Josiane Mothe *,†, Françoise Genova ‡

** IRIT, 118 route de Narbonne, 31062 Toulouse-Cedex 4, France hernandez@irit.fr , mothe@irit.fr*
† ERT34 - IUFM, 56 av. de l'URSS, 31078 Toulouse Cedex 4, France
‡ CDS, Centre de données de Strasbourg, Rue de l'université, Strasbourg

RÉSUMÉ. *Les techniques de recherche d'information s'appuient sur l'extraction de termes dans les documents, termes qui servent de base pour l'accès à ces documents. Ces techniques ont l'inconvénient de reposer sur des termes qui peuvent être ambigus et de ne pas prendre en compte les liens sémantiques qu'il existe entre les termes. Nous proposons dans cet article des fondations pour permettre une extraction plus riche sémantiquement en intégrant des connaissances issues de thésaurus et de corpus de domaine. Plus spécifiquement, nous proposons une méthodologie visant à transformer un thésaurus pré-existant en une ontologie légère de domaine qui sera utilisée pour indexer sémantiquement une collection de documents. Un corpus de référence est en outre utilisé pour compléter la connaissance représentée. Nous proposons également des techniques assurant cette transformation et une évaluation dans le domaine de l'astronomie.*

MOTS CLEFS : *Thesaurus, Ontologie, Création de ressources, langage d'indexation, Exploration de textes.*

ABSTRACT. *Information Retrieval techniques make use of terms that are automatically extracted from documents; these terms are used to give information access. In this paper we propose an approach to enrich semantically this extraction by adding knowledge from thesaurus. More specifically, the methodology we promote in this paper aims at transforming a thesaurus into a domain ontology which will then be used to semantically index documents (indexes are concepts rather than terms). We also propose techniques that implement this transformation as well as an evaluation in the field of the astronomy.*

KEYWORDS: *Thesaurus, ontology, resource acquisition, text mining.*

1. Introduction

La mise en œuvre de processus de gestion électronique de collections de documents a conduit à la création de nombreux thésaurus dont l'objectif est de contrôler la terminologie utilisée pour représenter de façon réduite les documents de la collection et de traduire en un langage plus strict (langage documentaire) le langage naturel utilisé dans les documents et dans les requêtes (Chaumier, 1988). Un thésaurus est fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels (AFNOR 1987). Les normes (ISO 2788 et ANSI Z39) ont permis d'uniformiser leur contenu en termes de relations entre unités lexicales : équivalence, hiérarchique et associative. Ce langage documentaire est ainsi utilisé pour indexer les documents de façon plus homogène. L'indexation s'appuyant sur un thésaurus est généralement réalisée de façon manuelle par des spécialistes de la documentation qui, à partir de leur expertise, choisissent à la lecture des documents les termes du thésaurus constituant l'index. Le même thésaurus est ensuite utilisé lors d'une recherche pour restreindre la portée d'une requête ou au contraire l'étendre, en fonction des besoins de l'utilisateur et du contenu de la collection. Cette approche est majoritairement utilisée dans les systèmes documentaires gérant des documents secondaires[1] électroniques pour permettre l'accès aux documents primaires qui eux sont au format papier.

Parallèlement, l'indexation automatique a été développée pour permettre la gestion de gros volumes de documents électroniques, souvent des documents primaires. Ces techniques reposent sur l'approche « sac de mots » : les textes sont analysés et les termes les plus représentatifs sont extraits des documents (Salton, 1971). Ce sont ces termes qui constituent alors le langage d'indexation et la base de la comparaison entre requêtes et documents. La pondération automatique des termes d'indexation (Roberston et Sparck-Jones, 1976), leur radicalisation (Porter, 1980), la reformulation de requête automatique par réinjection de pertinence (Harman, 1992) ou par ajout de termes co-occurrents (Qiu et Frei, 1993) sont des méthodes associées à l'indexation automatique et qui ont permis d'obtenir des performances intéressantes comme en attestent les programmes d'évaluation des systèmes de recherche d'information tels que TREC[2]. L'ensemble de ces approches fait l'hypothèse que les documents contiennent toute la connaissance nécessaire à leur indexation.

Afin de permettre une meilleure mise en correspondance des requêtes et des représentations des documents (index extraits) les systèmes de recherche d'information se sont également intéressés à l'utilisation des thésaurus. (Gauch et Smith, 1991) étend automatiquement ainsi les requêtes des utilisateurs en se basant

---

[1] Un document secondaire est un document issu de la description de documents primaires.

[2] TREC : Text REtrieval Conference http://trec.nist.gov

sur des relations entre termes issues d'un thésaurus. D'autres systèmes combinent l'utilisation de thésaurus à des mécanismes de classification (rattachement des documents au thésaurus) et de navigation. Le système Cat a Cone (Hearst, 1997) ou le système IRAIA (Englmeier et Mothe, 2003) s'appuient sur la structure hiérarchique des thésaurus pour permettre à l'utilisateur de naviguer au sein de sa structure et ainsi accéder aux documents associés aux termes. Cependant, le recours à un thésaurus soulève plusieurs problèmes : lorsqu'ils sont créés de façon manuelle, leur construction demande de lourds efforts, leur mise à jour est nécessaire dans le cas de domaines où la connaissance évolue, leur format n'est pas normalisé : fichiers ascii, html, bases de données co-existent ; enfin, les thésaurus possèdent un faible degré de formalisation puisque ils s'appuient sur la notion de termes plutôt que celle de concepts (Mizoguchi, 2004). Différentes solutions pour palier à ces inconvénients ont été proposées dans la littérature. La construction automatique de thésaurus peut ainsi faire appel à des techniques basées sur le calcul de corrélations entre termes (Tudhope et al., 2001), la classification automatique de termes (Bruandet et al, 1983), la classification de documents (Crouch et Yang, 1992) ou des approches prenant en compte des connaissances linguistiques (Grefenstette, 1992). D'autres part, les normes en cours d'élaboration dans le cadre du W3C comme SKOS Core[3] visent à faire migrer les thésaurus vers des ressources plus homogènes, sous un même format en se basant le langage OWL[4] et rendent ainsi ces ressources disponibles sur le web sémantique. Concernant la faible formalisation des thésaurus, la prise en compte des avancées en ingénierie des connaissances, en particulier au travers des ontologies est prometteuse. En effet, les thésaurus sont des collections de termes qui sont organisées suivant une ou plusieurs hiérarchies avec des relations entre termes. Etant utilisés par des documentalistes qui ont leur propre expertise sur le domaine de connaissance considéré, les thésaurus n'ont pas de niveau d'abstraction conceptuelle qui pourtant joue un rôle primordial dans la communication homme-machine (Soergel et al., 2004). Les ontologies permettent de reconsidérer ce problème puisqu'il s'agit d'une « *spécification formelle et explicite d'une conceptualisation partagée* » (Fensel, 1998). L'utilisation d'ontologies en Recherche d'Information permet de mettre en place un nouveau type d'indexation, appelée indexation sémantique. Contrairement à l'indexation par sac de mots, l'indexation sémantique repose sur l'idée selon laquelle le sens des informations textuelles (et des mots qui composent les documents ou les ressources) dépend des relations conceptuelles entre les objets du monde auxquels elles font référence plutôt que des relations linguistiques trouvées dans leur contenu (Haav et Lubi, 2001). L'indexation sémantique repose alors sur l'utilisation d'ontologies modélisant la conceptualisation des objets cités dans la collection à indexer. L'indexation sémantique, à partir de concepts plutôt qu'à partir de termes souvent ambigus, devient alors possible (Aussenac et Mothe, 2004). Cependant, l'élaboration d'une ontologie est coûteuse ; elle nécessite de nombreuses interventions manuelles. En effet, les techniques de construction d'ontologies de la

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

[4] http://www.w3.org/TR/owl-features/

littérature ne basent généralement pas l'élaboration de l'ontologie sur des connaissances préalables du domaine mais sur un corpus de référence qui est analysé (Aussenac et al, 2000).

Notre approche a pour originalité de 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 méthodologie pour transformer un thésaurus en ontologie légère de domaine pour l'indexation de corpus. Cette méthode vise à s'appliquer à n'importe quel thésaurus de domaine conçu en respectant les normes ISO 2788 et ANSI Z39. Ces thésaurus sont monolingues et ne sont pas organisés suivant des facettes. Nous présentons également différentes techniques permettant de mettre en œuvre la méthodologie proposée. Nous illustrons notre démarche à partir du thésaurus de l'astronomie IAU[5] ; les validations présentées s'appuient également sur ce thésaurus.

L'article est structuré de la façon suivante : la section 2 présente les différences fondamentales entre thésaurus et ontologies. La section 3 présente un état de l'art relatif à la conception d'ontologies. Dans les sections suivantes nous présentons notre approche. 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 (liens entre concepts) est construite. Enfin, la section 6 présente l'évaluation de notre approche dans le cadre de l'astronomie.

2. Thésaurus et ontologies

La principale distinction entre une ontologie et un thésaurus repose sur le degré d'engagement sémantique de ces deux représentations Le degré d'engagement sémantique correspond au niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique (Bachimont, 2000).

2.1. Normalisation du contenu d'un thésaurus

Les normes ISO 2788 et ANSI Z39[6] ont proposé les principes directeurs pour développer un thésaurus. Il s'agit d'une ressource terminologique dans laquelle les termes sont organisés suivant un nombre restreint de relations (Foskett, 1980) :

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

[6] http://www.techstreet.com/cgi-bin/pdf/free/228866/z39-19a.pdf

relations d'équivalence, hiérarchiques et associatives. Du point de vue de la représentation des connaissances, les thésaurus ont donc un faible degré de formalisation. 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 sémantique associée à un 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. Les relations « est plus spécifique » peuvent ainsi englober les relations « est une instance de » ou « est une partie de » (Fischer, 1998). La relation associative « est lié à » est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques (Tudhope et al., 2001). Par exemple, dans le thésaurus BIT[7] 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.

2.2. Ontologies

Alors qu'un thésaurus décrit un domaine en termes de catégories d'indexation, une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines. Elle définit le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine (Aussenac et Mothe, 2004).

2.2.1. Concepts

Un **concept** représente pour un objet matériel, une notion ou une idée (Uschold, 1995). Il est composé de trois parties : un ou plusieurs **termes**, une notion et un ensemble d'objets. La notion correspond à la sémantique du concept, elle est définie à travers ses relations aux autres concepts et ses attributs. La notion est appelée intention du concept. L'ensemble d'objets correspond aux objets définis par le concept, il est appelé extension du concept ; les objets sont les instances du concept. Le ou les termes permettent de désigner le concept. Ces termes sont aussi appelés **labels** de concept. Par exemple, le terme « siège » renvoie à un objet fabriqué, meuble disposé pour que l'on puisse s'y asseoir et à l'ensemble des objets ayant cette définition. Afin que les concepts soient reconnus de façon non ambiguë par la

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

machine, il est souhaitable qu'un concept soit référencé à partir de plusieurs termes, ce qui permet de gérer la synonymie et de les désambiguïser les uns par rapport aux autres (Gomez-Perez, 1996).

2.2.2. Relation entre concepts

Une relation sémantique R représente un type d'interaction entre les concepts d'un domaine c1, c2, .., cn.

La notion de subsomption (aussi appelée relation « est un », relation taxonomique ou relation de spécificité/généricité) est une relation binaire particulière qui implique l'engagement sémantique suivant (Guarino 2001) : un concept c1 subsume un concept c2 si toute relation sémantique de c1 est aussi relation sémantique de c2, en d'autres termes si le concept c2 est plus spécifique que le concept c1.

Les relations « associatives » sont des relations d'interaction entre deux concepts qui ne sont pas la relation de subsomption. La désignation «relation associative » est empruntée aux domaines de la bio-informatique (Zhang 2004), ce domaine ayant une utilisation équivalente des ontologies par l'indexation de publications et de comptes rendus biologiques. Elles correspondent à la notion de rôle en Logique de Description et permettent de typer les concepts reliés. Ces relations sont soit des propriétés entre concepts soit à des propriétés d'attribut dans le cas où elles associent un concept à un type de données. La sémantique qui leur est associée est référencée par un label. Elle peut également être précisée à partir de propriétés logiques associées à la relation telles que la transitivité, la symétrie, la fonctionnalité.

1.1.1. *2.2.3. Formalisation*

La structure d'une ontologie légère est définie par S := {C, R , A, T, , $\leq^C$, σ_R, σ_A } où :

- C, R, A, T, CAR_R sont des ensembles disjoints contenant les concepts, les relations associatives, les relations d'attribut, les types de données et les caractéristiques des relations associatives (synonymie, transitivité),
- $\leq^C$: C x C est un ordre partiel sur C, il définit la hiérarchie de concepts,

  ≤C (c1, c2) signifie que c1 subsume c2 (relation orientée)
- σ_R : R $\rightarrow$ C x C est la signature d'une relation associative,
- σ_A : A $\rightarrow$ C x T est la signature d'une relation d'attribut,

Le lexique d'une ontologie légère est un tuple L : $\{L^C, L^R, F, G\}$

- L^C L^R sont les ensembles disjoints des labels (termes) des concepts et des relations,
- F, G sont deux relations appelées référence,

$F : L^C \rightarrow C$ pour les concepts et $G : L^R \rightarrow R$ pour les relations

- Pour $l \in L^C$, $F(l)=\{c / c \in C\}$
- Pour $c \in C$, $F^{-1}(c)=\{l / l \in L^C\}$
- Pour $l \in L^R$, $G(l)=\{r/r \in R\}$
- Pour $r \in R$, $G^{-1}(r)=\{l / l \in L^R\}$

Ces relations permettent d'accéder aux concepts et relations désignés par un terme et réciproquement. Notons qu'un concept peut être désigné par différents termes et qu'un terme, dans le cas où il est ambigu, peut référencer différents concepts. Nous utiliserons cette formalisation dans l'écriture des règles de notre méthode.

D'autre part, le langage OWL[8] est un langage permettant de représenter une ontologie ; c'est celui que nous avons choisi dans la mesure où il a été retenu par le W3C.

3. Méthodes de construction d'ontologies : état de l'art

La conception d'ontologies est une tâche difficile nécessitant la mise en place de procédés élaborés afin d'extraire la connaissance d'un domaine, manipulable par les systèmes informatiques et interprétable par les êtres humains. Deux types de conception existent : la conception entièrement manuelle et la conception reposant sur des apprentissages. Plusieurs principes et méthodologies ont été définis pour faciliter la construction manuelle. Ces principes se basent sur des fondements philosophiques et suivent des procédés de modélisation collaboratifs. Ils mènent à la conception d'ontologies dites légères et d'ontologies dites lourdes (ces ontologies se distinguent par la présence ou non d'axiomes). Cependant, ce procédé de génération est très coûteux en temps et pose surtout des problèmes de maintenance et de mise à jour (Ding, 2002). La conception automatique d'ontologies commence à émerger comme un sous-domaine de l'ingénierie des connaissances. Face à la masse croissante de documents présents sur le Web et aux avancées technologiques dans le domaine de la recherche d'information, de l'apprentissage automatique et du traitement automatique des langues, de nouveaux travaux portent sur la recherche de procédés plus automatiques de génération d'ontologies. Ces mécanismes mènent généralement à la conception d'ontologies dites légères. Dans (Maedche et Staab, 2001), différents types d'approches sont distingués en fonction du support sur lequel elles se basent : à partir de textes, de dictionnaires ou thésaurus, de bases de connaissance, de schémas semi-structurés et de schémas relationnels. Dans le cadre de nos travaux, nous nous sommes plus particulièrement intéressés aux approches

[8] http://www.w3.org/TR/owl-features/

basées sur les textes dans la mesure où elles ouvrent des perspectives, selon nous, à un enrichissement incrémental de ressources.

Afin de transformer un thésaurus en une ontologie, les méthodes développées visent à capturer la sémantique informelle du thésaurus soit manuellement (Wielinga et al. 2001), soit à partir de patrons syntaxiques (Soergel et al. 2004), soit à partir d'inférences (Hahn et Schulz, 2004). Le traitement entièrement manuel n'est envisageable que dans des cas très limités compte tenu de son coût. Un traitement semi-automatique paraît donc plus adapté. Une première contribution (Soergel et al. 2004) vise à l'aide à cette transformation. Cependant, le travail manuel demandé reste important puisque l'expert doit proposer des patrons à partir de l'analyse de chaque couple de termes. Notre contribution vise à limiter le travail de l'expert en lui demandant de proposer des relations entre concepts d'un haut niveau hiérarchique plutôt qu'entre couples de termes. De plus, une limite des approches de la littérature est qu'elles ne construisent l'ontologie qu'à partir de la connaissance contenue dans un thésaurus. Cette connaissance ne reflète pas nécessairement l'évolution de la connaissance du domaine. Nous proposons de transformer un thésaurus en s'appuyant sur la connaissance qu'il représente et sur les informations contenues dans un corpus documentaire du domaine.

Concernant la conception d'ontologies à partir de textes, différents outils de conception ont été développés[9]. Chacun d'entre eux présente des fonctionnalités différentes et a permis l'élaboration de nombreuses ontologies. Text-To-Onto, développée à l'Institut AIFB de l'Université de Karlsruhe, est une application d'extraction d'ontologies à partir de corpus ou de documents Web qui permet également la réutilisation d'ontologies existantes (Maedche et Staab, 2001). Text-To-Onto est intégrée à la plate-forme logicielle KAON qui permet l'édition et la maintenance d'ontologies (Bozsak et al., 2002). KAON utilise le langage de représentation RDFS et est orientée vers l'utilisation des ontologies sur le Web, l'application KAON Portal permettant la recherche et le parcours d'ontologies via un navigateur Web. OntoBuilder, développée au Technion d'Haifa, permet de bâtir une ontologie à partir de ressources Web (Gal et al, 2004). L'extraction de l'ontologie à partir de fichiers XML est suivie d'une phase de raffinage guidée par l'utilisateur. Onto-Builder autorise aussi la fusion d'ontologies extraites de différents sites Web.

La méthodologie TERMINAE de (Aussenac et al., 2000) propose une approche pour sélectionner les concepts, leurs propriétés, les relations et leur regroupement. Cette méthodologie a été développée dans le laboratoire et est une composante de la plate forme RFIEC[10]. Elle repose sur l'utilisation d'outils de traitement automatique des langues analysant les termes de textes et les relations lexicales. Les termes sont

---

[9] Pour une étude comparative détaillée des outils de conception d'ontologies, nous renvoyons le lecteur aux travaux de Su et Ilebrekke (Su et Ilebrekke, 2002) et de Sure et Corcho (Sure et Corcho, 2003)

[10] RFIEC est une plateforme regroupant un ensemble de résultats associé aux compétences locales d'équipes travaillant autour de l'analyse et la représentation de textes (outils, méthodologies, corpus, ressources linguistiques)

regroupés suivant leur contexte et facilitent la création de concepts et de relations sémantiques. Les concepts et relations sont ensuite formalisés dans un modèle. TERMINAE a l'avantage de répondre à certaines questions et d'axer le choix des concepts et des relations de l'ontologie sur l'extraction de termes d'un corpus de référence. Notre méthodologie d'élaboration d'ontologies à partir de textes prolonge TERMINAE en intégrant les ressources terminologiques que sont les thésaurus.

La méthodologie que nous proposons vise à permettre l'élaboration d'une ontologie légère de domaine pour la RI 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 d'information actuelle relative au domaine, la méthode repose sur l'analyse de documents textuels selon TERMINAE. Associée à cette méthodologie, nous proposons une implantation dans laquelle nous minimisons le travail manuel.

La première étape 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 la RI, nous avons identifié les besoins suivants :

- L'identification des termes du domaine et de leurs variantes lexicales : l'ontologie doit permettre de représenter au mieux le contenu des granules dans la phase d'indexation sémantique et celui des besoins d'information dans la phase de recherche. Ces termes doivent donc correspondre à une couverture minimale nécessaire pour l'utilisation en RI ; ils sont extraits de façon automatique du thesaurus et des textes.
- Le regroupement de ces termes en concepts afin de déterminer les objets et notions référencés dans les documents ou les requêtes. Ce regroupement repose sur une approche automatique à partir de connaissances extraites de ressources.
- La structuration des concepts à partir de relations taxonomiques et associatives afin de permettre une indexation sémantique de qualité ou une reformulation sémantique de la requête. Cette structuration repose sur une approche automatisée. Les experts n'intervenant que dans la validation des concepts abstraits.
- La formalisation de l'ontologie dans un langage interprétable par le SRI afin qu'il soit capable de la manipuler. Cette formalisation utilise sur OWL.

La deuxième étape repose sur le choix du corpus de référence à partir duquel l'ontologie est construite de façon automatisée. Ce choix est un paramètre déterminant de l'élaboration de l'ontologie. Dans notre approche, le corpus est extrait de corpus existants et des experts doivent s'assurer de la couverture du domaine sur une période représentative. Des résumés d'articles publiés dans des revues du domaine permettent d'obtenir ce type d'information. Les articles complets pourraient être utilisés mais l'avantage des résumés lié à la présence d'information synthétique.

La troisième étape est celle de l'étude des ressources : thésaurus et corpus. Cette étape vise à extraire 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. Cette étape intègre la connaissance représentée dans le thésaurus : les termes présents dans le thésaurus sont regroupés à partir des relations de celui-ci. L'étude linguistique du corpus de référence permet également d'extraire des termes du domaine et les relations entre termes qui ne sont pas explicitées dans le thésaurus. Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique SYNTEX[11] (Bourigalut et Fabre, 2000). Cet analyseur a l'avantage de reposer 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). Il est cependant nécessaire 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 répond à ce problème.

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. Dans notre approche, nous avons choisi le langage OWL. Ce langage 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 en RI.

4. 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. Chaque concept possèdera alors un ensemble de label correspondant aux termes du langage utilisés pour représenter ce concept. Dans un processus d'indexation de documents ou de mise en correspondance entre requêtes et documents, cela revient à limiter le nombre de variantes lexicales à considérer puisque le niveau concept seul est utilisé.

[11] SYNTEX a été développé par D. Bourigault, membre de l'équipe ERSS et partenaire de la plateforme RFIEC.

4.1. Utilisation des 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 des termes préférés et de l'ensemble des termes auxquels ils sont liés par les relations UP et UPD (Règle R1).

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)**

4.2. 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. La figure 1 schématise plusieurs exemples de groupements. Pour faciliter la lisibilité, les termes préférés sont en gras majuscules.

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)**

Extrait d'un thésaurus (IAU)

ELLIPSOIDAL VARIABLE STARS UPD photometric binary stars
ellipsoidal binary stars UP **ELLIPSOIDAL VARIABLE STARS**
→Exemple de termes regroupés par R1
ELLIPSOIDAL VARIABLE STARS *ellipsoidal binary stars photometric binary stars*

ZENITH TUBES UP zenith telescopes
ZENITH TELESCOPES UP **photographic zenith tubes**
→Exemple de termes regroupés par R2

FIG.1 - *Exemples de groupements des termes du thésaurus*

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

4.3. Identifiant du concept

L'identifiant d'un concept est déterminé par le terme préféré à l'origine du groupement. Ce choix permet de garder un lien entre l'ontologie et le thésaurus. Les identifiants des concepts correspondent ainsi à des entrées du thésaurus. Un terme peut être polysémique (label de plusieurs concepts) s'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)

5. Construction de la structure de l'ontologie

La structure de l'ontologie définit les relations entre concepts extraits comme décrit précédemment. La structure comprend des relations taxonomiques de type « est un » et des relations associatives.

5.1. Construction de la hiérarchie de concepts

Certains liens hiérarchiques entre concepts sont directement issus des liens explicitement 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 2 schématise ces différents mécanismes.

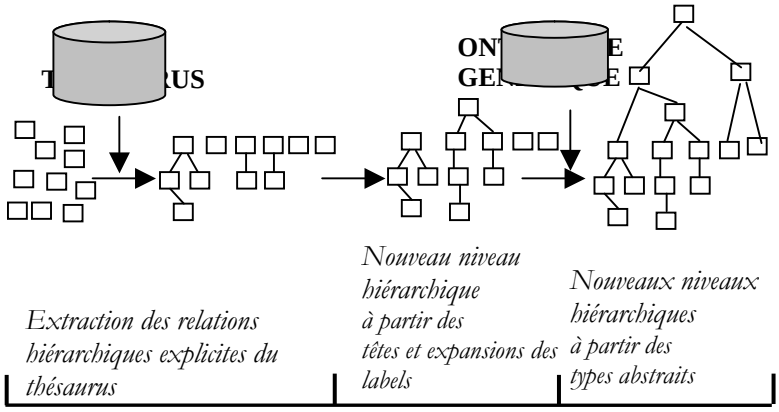

FIG. 2 - *Mécanisme de construction de la hiérarchie de concepts*

5.1.1.Extraction 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 l'ensemble des 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. Beaucoup de thésaurus de domaine prennent la peine de considérer les relations « est plus spécifique que » et « plus générique que » de façon stricte (ce qui est le cas du thésaurus IAU).

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)

5.1.2. 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 : A,B,C étant des concepts, si A « est une sous classe de » B et B « est une sous classe de » C, alors A « est une sous classe de » C,. Cette dernière relation n'a donc pas besoin d'être présente dans 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 ≠chem2
=> suppression de l'arc à l'origine du chemin le plus court
(R5)

5.1.3. Nouveaux niveaux hiérarchiques

Une des lacunes des thésaurus est que leur plus haut niveau hiérarchique contient généralement un très grand nombre de termes (Soergel et al., 2004). Ceci s'explique par le fait que les thésaurus ne définissent pas de catégories génériques permettant de répertorier l'ensemble des termes du domaine. Cela implique que la même lacune

se retrouve dans l'ontologie obtenue par la transformation d'un thésaurus à partir des règles précédentes. 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 élevés et la définition de concepts génériques (ou types abstraits) permettant de caractériser les concepts. Un concept générique n'admet pas d'instance. Nous proposons une approche automatisée pour créer ces niveaux hiérarchiques supplémentaire.

1ier niveau de généralisation : tête et expansion des syntagmes

Pour créer un premier niveau de généralisation, les concepts sont regroupés à partir de la tête des termes de leur label. Cette approche est également suivie dans OntoLearn (Velardi et al., 2001) 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 3). Si ce concept n'existe pas dans l'ontologie, il est créé et appartient au nouveau niveau 0 de l'ontologie (règle R7 et figure 4).

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)

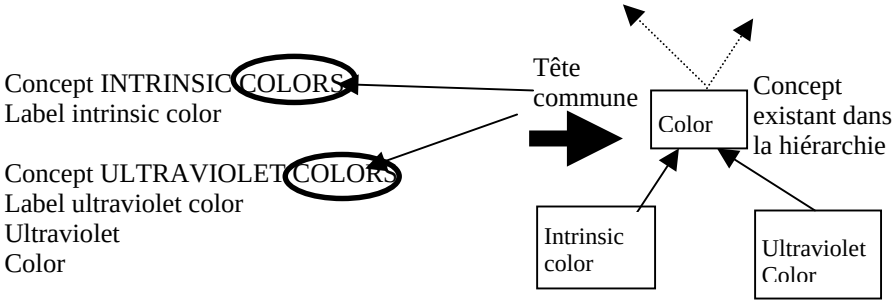

FIG. 3. - *Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie*

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)

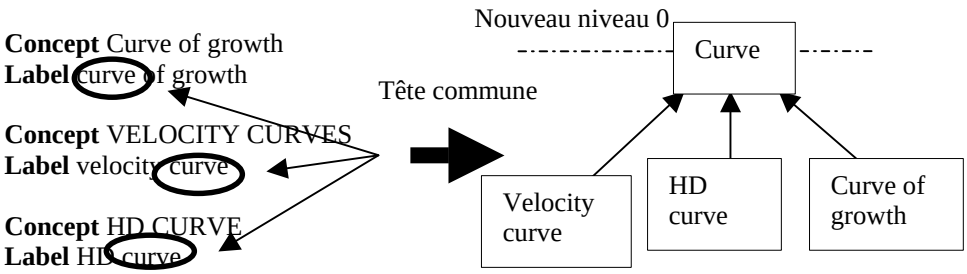

FIG. 4.- *Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie*

2ième 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 ⇔ sw avec sw∈{Synsets_WordNet}
=> c « est sous classe de » ta
Avec ta (type abstrait) est le plus spécifique hyperonyme de sw
(R8)

- *Définition des types abstraits*

La définition de types abstraits de haut niveau implique l'intervention humaine. Alternativement, il est possible d'avoir recourt à une ontologie de haut niveau qui contient ces concepts très généraux, interdisciplinaires. Nous avons choisi d'utiliser WordNet pour sa disponibilité, du moins en langue anglaise.

Les concepts le plus généraux de la nouvelle ontologie sont mis en correspondance avec les concepts de l'ontologie générique. Les types abstraits sont alors définis à partir des concepts les plus génériques associés aux concepts communs.

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. 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. Si un de ces termes est retrouvé, le Synset candidat est automatiquement choisi.

(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 (échec de la désambiguisation).

Concernant l'identification des types, les Synsets les plus génériques (i.e. les plus lointains 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 comme 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 5 présente des exemples de types abstraits extraits dans 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*

FIG. 5.- *Extrait du nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie*

5.2. Extraction 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.

5.2.1. 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 »

$$=> G^{-1}(r') \in G^{-1}(r)$$

(R9)

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.

Définition de relations entre types

La définition des relations sémantiques est réalisée entre chaque paire de types abstraits. 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, qui relient des concepts de même type. Pour ce type de relation, la proposition « partie de » est ajoutée. 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**
Property	*Influences / Is influenced by* *Determined by /Determines* *Exclude* *Has part / Is part*	*Is a property of* *induces*	*Is a property of* *induces*	*Is studied by*
Instru-mentation	Makes Observes	*Observes* *Measures*	*Observes* *Measures*	*Is Used to studied*

TAB. 1 - ***Extrait de la matrice des relations entre types abstraits***

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 et al., 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.

5.2.2. Extraction 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 choisis par des experts du domaine, il n'y a pas vraiment de limitation dans le nombre de documents dans la mesure où leur analyse est automatique (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 (Bourigault, 2002). 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. Nous proposons 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 $\sigma_{R Onto}$: $R_{Onto} \rightarrow C \times C$ 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. Evaluation

Nous présentons dans cette section un retour d'expérience sur l'application de la méthode de transformation d'un thésaurus en ontologie légère ; cette application s'appuie sur le thésaurus IAU. Le thésaurus IAU a été conçu dans l'objectif de

standardiser la terminologie du domaine de l'astronomie. Son utilisation est destinée à aider les documentalistes dans la désambiguïsation des mots clés choisis pour indexer les catalogues et les publications scientifiques du domaine. Sa conception, demandée par l'Union Internationale de l'Astronomie en 1984, a été terminée en 1995. Sa transformation en ontologie légère a été réalisée dans le cadre du projet Masses de Données en Astronomie[12]. Il s'inscrit dans le cadre de l'élaboration d'un observatoire virtuel. Il vise à proposer des solutions quant à l'utilisation scientifique optimale des informations du domaine de l'astronomie, notamment par l'indexation sémantique des documents numériques textuels du domaine.

La transformation du thésaurus IAU a été réalisée à partir de l'ensemble des règles de transformation que nous avons définies. Deux corpus du domaine de l'astronomie ont été utilisés. Les documents qu'ils contiennent sont des résumés d'articles publiés dans la revue internationale Astronomy and Astrophysics (A&A). Ces documents sont en langue anglaise. Le premier corpus est composé d'articles publiés en 1995. Il vise à aider à la capture de la connaissance non représentée dans le corpus au moment de sa création. Le deuxième corpus est constitué d'articles publiés en 2002. Ce corpus a été choisi pour permettre la mise à jour des connaissances du domaine à partir de documents récents (non présenté dans cet article). Les deux corpus ont été validés par des experts du domaine pour décrire les connaissances à représenter dans l'ontologie.

Le protocole choisi pour évaluer ces règles et les résultats de leur évaluation sont présentés dans cette section.

6.1. Protocole

Le protocole d'évaluation défini consiste à présenter les résultats obtenus par les différentes règles sur un des échantillons du thésaurus et à les faire valider par deux astronomes qui acceptent ou rejettent les propositions qu'elles permettent d'obtenir. Pour chacune des règles, l'ensemble des propositions est présenté à l'expert du domaine en respectant le même format. Les résultats sont ensuite dépouillés à partir des fichiers annotés par les experts.

L'extraction des concepts et des relations hiérarchiques à partir du thésaurus correspondent à des règles simples (R1 à R5). L'évaluation de ces éléments revient à évaluer la pertinence du thésaurus initial. Les résultats que nous avons obtenus montrent que les experts étaient dans l'ensemble d'accord avec l'exactitude de la connaissance initiale contenue dans ce thésaurus.

6.2. Types abstraits

Au niveau le plus générique de l'ontologie créée après application des règles R1 à R5 de la transformation, 1132 concepts ont été définis. Les différentes étapes de

[12] http://cdsweb.u-strasbg.fr/MDA/mda.html

réduction de ce nombre par recherche de concepts généralisant nous conduit après application des règles R6, R7 et R8 à obtenir les types abstraits des concepts du plus haut niveau de généralisation à partir d'une ontologie générique. Dans notre cas, le choix de cette ontologie s'est porté sur WordNet car pour 72 % des concepts du plus haut niveau, au moins un de leur label est défini dans la ressource. L'étape de désambiguïsation proposée permet d'identifier pour 65% d'entre eux un seul Synset auquel ils sont associés. A partir de ces Synsets généralisant les Synsets associés aux concepts de l'ontologie, 19 types abstraits ont été proposés aux astronomes. Ces types abstraits ont été présentés à partir de la définition des Synsets donnée dans WordNet et des concepts de l'ontologie desquels ils étaient extraits ; 14 ont été retenus par les astronomes. Des échantillons des concepts rattachés à chacun des types ont été analysés. Ces échantillons sont choisis aléatoirement et représentent 80 % des concepts associés à chaque type. La pertinence des rattachements est présentée dans le tableau 2.

Type abstrait	**Nombre de concepts évalués**	**% de concepts pour lesquels le lien au type est correct**
PROPERTY	53	75%
PHENOMENON	68	67%
EVENT	14	42%
SCIENCE	30	93%
INSTRUMENTATION	13	100%
SUBSTANCE	4	100%
RELATION	19	100%
ANGLE	5	100%
PLANE	4	100%
REGION	15	100%
NATURAL_OBJECT	10	100%
ARTEFACT	34	85%

Tab. 2. - ***Proportion des concepts correctement rattachés à un type abstrait***

Le résultat des rattachements des concepts est globalement très positif. Il permet en moyenne d'associer les concepts à un type abstrait avec 89% de précision. Les rattachements au type *event* sont les moins pertinents. Ceci s'explique par le fait que les astronomes considèrent que les hyperonymes associés dans WordNet aux Synsets définis à partir des labels de ces concepts de l'ontologie descendant de ce type abstrait ne sont pas adaptés pour l'astronomie et correspondent plutôt à des concepts de la physique.

6.3 Spécification des relations associatives entre concepts

Les types abstraits obtenus sont utilisés pour préciser les relations entre concepts de l'ontologie.

6.3.1. Relations au niveau des types abstraits

La première étape dans la spécification des labels des relations entre concepts est la définition des relations entre types abstraits.. Elle est réalisée par les astronomes sur la base des termes détectés dans le contexte d'apparition des labels des concepts dans le corpus de référence. Les termes apparaissant fréquemment dans le contexte sont regroupés en fonction du type abstrait donc descendent les concepts à côté desquels ils apparaissent. Les astronomes se sont plutôt inspirés du contexte représenté par les verbes. Cette remarque confirme l'intuition de certains travaux de la littérature qui font reposer la spécification des relations associatives entre concepts par les verbes du corpus. La spécification des relations entre types abstraits a nécessité deux heures de travail pour les experts.

6.3.2. Désambiguïsation des relations « est lié à »

Des relations entre types ont été proposées pour spécifier la nature des relations vagues entre concepts extraits à partir des relations « est lié à » du thésaurus. Cette étape correspond à la règle R9 décrite dans la section 5.2.1. Les résultats obtenus pour ces relations définies pour les concepts descendant des types abstraits *instrumentation* et *property* sont présentés dans le tableau 3.

	Nombre de relations vagues évaluées	**Nombre de relations incorrectement labellisées**
Concepts descendant du type abstrait property	34	5
Concepts descendant du type abstrait instrumentation	15	3

TAB. 3 - *Résultat de la désambiguïsation des relations « est lié à » entre concepts extraits du thésaurus*

Les résultats des expérimentations montrent que l'utilisation de la matrice de relations sémantiques entre concepts s'applique très concrètement à la désambiguïsation des relations vagues du thésaurus. Pour les relations qui n'étaient pas correctement labellisées, les astronomes ont proposé deux nouveaux labels qui ont été intégrés à la matrice.

6.3.3. Détection de nouvelles relations

Les relations entre types sont également utilisées pour caractériser de nouvelles relations entre les concepts existant dans l'ontologie. La règle R10 spécifie cette étape. Elle consiste à prendre en compte le contexte dans le corpus de référence des différents labels descendant des types abstraits et à proposer une nouvelle relation

entre deux concepts dans le cas où leurs labels apparaissent dans le contexte de l'un et de l'autre. La relation est alors labellisée à partir des types abstraits desquels descendent les deux concepts par la matrice précédemment réalisée. Deux approches ont été proposées pour extraire le contexte d'un label et pour mettre en place cette règle.

La première repose sur l'analyse des termes avec lesquels un label co-occurre. Les termes qu'il régit ou par lesquels il est régi sont alors étudiés. Pour évaluer cette approche, nous avons analysé 50% des relations ainsi extraites du corpus de référence pour les types abstraits *instrument* et *property (cf* tableau 4).

	Nombre de relations proposées	**Nombres de relations proposées incorrectes**	**Nombres de relations dont le label proposé est incorrect**
Concepts descendant du type abstrait property	47	3	2
Concepts descendant du type abstrait Instrumentation	27	2	8

TAB.4 - *Résultat de l'analyse des nouvelles relations entre concepts proposées à partir du contexte de leur label dans le corpus*

Les résultats de l'évaluation des nouvelles relations proposées entre concepts à partir du contexte de leurs labels montrent qu'une forte proportion des relations est correcte. Les labels proposés pour ces relations sur la base de la matrice des types sont pour la plupart également validés. Notons, cependant, que les astronomes ont jugé que certaines relations ne s'appliquaient pas uniquement aux concepts au niveau desquels elles étaient décelées mais pouvaient être généralisées à certains de leurs concepts pères. Ces relations sont d'ailleurs dans quelques cas décelées pour leurs pères. Cette remarque a mené à une nouvelle proposition pour l'implantation de cette étape. Elle consiste à analyser les nouvelles relations entre concepts par leur niveau hiérarchique dans l'ontologie. Les relations détectées sont ensuite héritées par les concepts fils. Pour chaque concept, seules les relations qu'aucun des ancêtres ne possède sont évaluées.

La deuxième méthode d'extraction du contexte d'un label repose sur l'analyse distributionnelle réalisée par le module UPERY de Syntex. Ce type d'analyse consiste à rapprocher des labels en fonction de la ressemblance de leur contexte. Ils sont rapprochés s'ils sont formés dans le corpus autour des mêmes relations syntaxiques et des mêmes têtes et queues. Un coefficient de proximité entre termes est défini dans le module. Ce coefficient tient compte de la productivité des contextes partagés par les termes, la productivité correspondant aux nombres de termes qui partagent le contexte. Le coefficient de proximité repose sur le principe

suivant : si un contexte partagé par deux termes est très productif, sa contribution au rapprochement des deux termes est a priori plus faible que celle d'un contexte peu productif. Nous avons donc sélectionné, pour évaluer notre approche, les termes rapprochés par un coefficient de proximité inférieur ou égal à 80% de la proximité maximale. Le tableau 5 présente les résultats de l'évaluation des relations ainsi décelées pour les concepts descendants des types *property*.

	Nombre de relations proposées	**Nombres de relations proposées qui ne sont pas correctes**	**Nombres de relations dont le label proposé est incorrect**
Concepts descendant du type abstrait property	48	40	3

TAB.5 - *Résultat de l'analyse des nouvelles relations entre concepts proposées à partir du module UPERY*

Les résultats de cette évaluation montrent que les relations décelées sont pour la plupart erronées. Le rapprochement des termes en fonction des contextes qu'ils partagent ne permet pas de déterminer des relations entre concepts.

7. Conclusion

Dans cet article, nous avons proposé une méthodologie permettant d'augmenter la représentation sémantique d'un domaine ainsi que sa formalisation. Nous nous appuyons pour cela sur un thésaurus du domaine, sur un corpus de référence et sur une formalisation sous la forme d'une ontologie. Cette méthodologie est accompagnée de méthodes permettant son implantation. Ces méthodes reposent sur des outils qui mettent en œuvre des compétences inter-disciplinaires comme par exemple la gestion de connaissances avec TERMINAE, la linguistique avec SYNTEX. Un point fort de notre proposition concerne le fait que l'ensemble fait appel soit à des outils développés localement et intégrés dans notre plateforme locale RFIEC, soit des ressources disponibles de façon internationale (TreeTagger sur lequel repose SYNTEX, WordNet).

Notre approche vise à minimiser le travail des experts qui sont généralement fortement sollicités. Le procédé de transformation d'un thésaurus en ontologie légère repose sur quatre étapes principales : l'extraction d'informations du corpus, l'identification des concepts issus du thésaurus, la construction de la structure de l'ontologie (hiérarchie de concepts et relations associatives entre concepts). Les procédés sont simples à mettre en œuvre et permettent d'extraire automatiquement

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. L'expert est moins sollicité que dans les approches proposées dans (Soergel et al., 2004) et (Wielinga et al., 2001) car son travail consiste uniquement à valider les propositions. Contrairement aux approches présentées dans la littérature, le procédé mis en place vise à transformer le thésaurus.

Une contribution importante présentée dans cet article 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 tenu 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 ».

L'évaluation de la méthode de transformation de thésaurus en ontologie sur le domaine de l'astronomie a montré son intérêt. Elle permet de déterminer un ensemble de concepts ainsi que leurs labels pertinents pour le domaine. De plus, elle extrait efficacement des types abstraits qui sont associés aux concepts les plus génériques de l'ontologie. Ces types abstraits structurent l'ontologie et facilitent la désambiguïsation et la détection de relations associatives entre concepts.

A la suite de cette évaluation, plusieurs perspectives sont envisagées. L'intégration de nouvelles connaissances dans l'ontologie (ajout de termes, de relations entre concepts) est un des aspects importants. Cette optique est primordiale car la date de création des thésaurus remonte souvent à plusieurs dizaines d'années et la connaissance d'un domaine évolue rapidement. Elle permet également d'avoir recourt à une mise à jour incrémentale de l'ontologie (nouveaux corpus à indexer par exemple). La méthode proposée ici a été complétée en ce sens (Chrisment et al. 2006). Une perspective intéressante concerne la définition d'une méthode pour associer aux types abstraits les concepts de l'ontologie qui ne peuvent être associés aux Synsets de WordNet (soit parce que leurs labels ne figurent pas dans cette ontologie générique, soit parce que la désambiguïsation des Synsets associés n'est pas possible). Nous envisageons pour cela de prendre en compte l'ontologie générique DOLCE qui nous permettra d'apporter un degré de formalisation plus important à notre ontologie comme il est montré dans (Fortier et Kassel, 2004). Une méthode devra également permettre d'aider l'expert dans le choix des relations entre concepts proposées si celles-ci peuvent avoir plusieurs labels.

Remerciements

Les travaux présentés dans ce papier ont bénéficié du cadre du projet Masse de Données en Astronomie supporté par le ministère délégué à la Recherche et aux Nouvelles Technologies. Nous tenons à remercier particulièrement Pascal Dubois, Andrea Preite Martinez, astronomes du CDS qui ont évalué nos propositions. Nous remercions également Didier Bourigault pour l'utilisation des logiciels qu'il a conçus .

8. Références

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

N. Aussenac-Gilles, J. Mothe, Ontologies as Background Knowledge to Explore Document Collections, Actes de RIAO, pages 129-142, 2004.

B. Bachimont, Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances, In Ingénierie des connaissances : évolutions récentes et nouveaux défis, pp 305–323, Eyrolles, 2000.

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), pages 75-84, 2002.

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

MF. Bruandet, Y. Chiaramella, D. Kerkouba, Etude de NEF, Projet CONCERTO, 1983.

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

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

Y. Ding, S. Foo, Ontology Research and Development: Part 1 – A Review of Ontology Generation, Journal of Information Science 28(2), 2002.

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

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

M. Fernandez, A. Gómez-Pérez, N. Juristo, METHONTOLOGY: from ontological art towards ontological engineering, Actes de AAAI, 1997.

D. H. Fischer, From Thesauri towards Ontologies?, dans: el Hadi, Maniez & Pollitt (Eds.): Structures and Relations in Knowledge Organization, dans 5th Int. ISKO Conference, pages, 18-30, 1998.

J.-Y. Fortier & G. Kassel, Managing Knowledge at the Information Level: an Ontological Approach. In Proceedings of the ECAI'2004 Workshop on Knowledge Management and Organizational Memories, August 22-27, Valencia (Spain), p. 39-45, 2004.

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

A. Gal, G. Modica, H.M. Jamil, OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources, In Proceedings of the 20th International Conference on Data Engineering, IEEE Computer Society, 2004.

S. Gauch et J.B. Smith, Search Improvement via Automatic Query Reformulation, ACM Transactions on Information Systems, 9(3), pages 249-280, 1991.

G. Grefenstette, Use of syntactic context to produce term association lists for retrieval, Conference on Research and Development in Information Retrieval (SIGIR), pages 89-97, 1992.

A. Gómez-Pérez, M. Fernandez, A.J. de Vicente, Towards a Method to Conceptualize Domain Ontologies, In Proceedings of the European Conference on Artificial Intelligence (ECAI'96), pp 41–52, 1996.

N. Guarino, C. Welty, Identity and Subsumption, In The Semantics of Relationships: an Interdisciplinary Perspective, R. Green, C.A. Bean, S. Hyon Myseng (Eds), Kluwer, pp 111-126, 2001.

Y. Guo, H. Harkema, R. Gaizauskas. Sheffield University and the TREC 2004 Genomics Track : Query Expansion Using Synonymous terms, 2004.

U. Hahn, S. Schulz, Building a Very Large Ontology from Medical Thesauri, Handbook on Ontologies, S. Staab, R. Stuber (Eds.) pp 133-150, 2004.

D. Harman, Relevance feedback revisited, Conference on Research and Development in Information Retrieval (SIGIR), pages 1-10, 1992.

H.M. Haav et T.L. Lubi, 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, 2001.

M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, Conference on Research and Development in Information Retrieval (SIGIR), pages 246-257, 1997.

C. Chrisment, N. Hernandez, G. Hubert, J. Mothe, Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents, soumis à la revue 3I, numéro spécial Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance, 2006.

A. Maedche, S. Staab. Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2), 2001.

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

R. Mizoguchi, Le rôle de l'ingénierie ontologique dans le domaine des EIAH, Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation, Vol. 11, 2004.

MF. Porter, An Algorithm for Suffixing Stripping, Program, 1(3), pages 130-137, 1980.

Y. Qiu et H.P. Frei, Concept based Query Expension, Conference on Research and Development in Information Retrieval (SIGIR), pages 160-169, 1993.

SE. Robertson, et K. Sparck-Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27 (3), pages 129-146, 1976.

G. Salton, the SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton Ed., Prentice Hall Inc., 1971.

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

X. Su et L. Ilebrekke, A comparative study of ontology languages and tools , Conference on Advanced Information System Engineering (CAiSE' 02). Toronto, Canada, 2002.

Y. Sure, Ó. Corcho, EON2003, Evaluation of Ontology-based Tools, International Workshop on Evaluation of Ontology-based Tools held at the 2nd International Semantic Web Conference (ISWC), USA CEUR-WS.org, 2003.

M. Uschold, M. King, Towards a Methodology for Building Ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995), 1995.

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

C. J. van Rijsbergen, D. J. Harper, and F. Porter, M. The selection of good search terms. Information Processing & Management, 17(2), pages 77-91, 1981.

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

B. Wielinga, G. Schreiber, J. Wielemaker, J.A.C. Sandberg, From thesaurus to ontology, In Proceedings of the International Conference on Knowledge Capture, 2001.