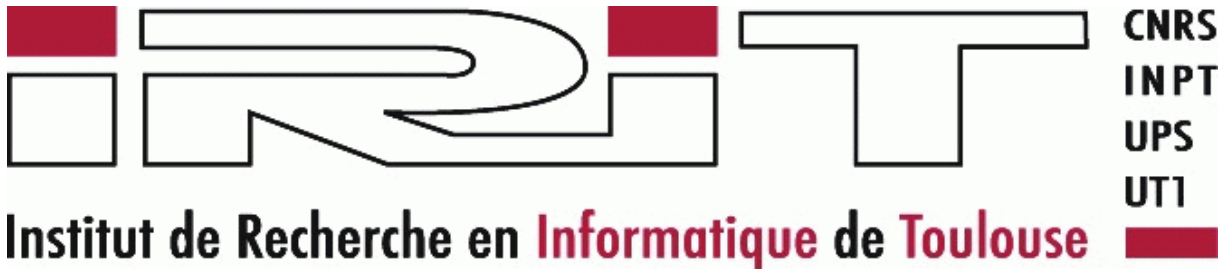
Universités Paul Sabatier Toulouse III et des
Sciences Sociales de Toulouse 1



RAPPORT INTERNE IRIT

IRIT/RR—2006-04--FR

FEVRIER 2006

NATHALIE HERNANDEZ, JOSIANE MOTHE

TtoO: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence

TtoO: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence

Nathalie Hernandez (hernandez@irit.fr), Josiane Mothe (mothe@irit.fr)

IRIT, 118 route de Narbonne, 31062 Toulouse-Cedex 4, France

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. Nous proposons dans ce rapport des fondations pour permettre une extraction plus riche sémantiquement en intégrant des connaissances issues de thésuaurs. Plus spécifiquement, nous proposons une méthodologie visant à transformer un thésaurus pré-existant en une ontologie de domaine qui sera utilisée pour indexer sémantiquement, c'est-à-dire à partir de concepts plutôt que de termes, 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.

Nos propositions s'appuient d'une part sur la connaissance présente dans un thésaurus et sur celle que nous extrayons automatiquement d'un corpus de référence. Ainsi, certaines relations entre termes présentes dans le thésaurus sont directement exploitées pour formaliser la connaissance sous forme d'ontologie (relations « Utiliser plutôt », « Utiliser pour désigner », « est plus générique que »). D'autres connaissances sont directement extraites de l'analyse du corpus (nouveaux labels et liens hiérarchiques en particulier). Enfin, une ressource générique est utilisée pour définir des types abstraits permettant de hiérarchiser les concepts de haut niveau. Ces différentes ressources sont utilisées de façon complémentaire pour terminer l'enrichissement de l'ontologie (désambiguïsation de la relation « est lié à » issue du thesaurus). Les résultats de l'évaluation que nous avons mise en place sur un échantillon du thesaurus IAU en astronomie montrent que l'approche est satisfaisante puisque dans 89% des cas, le type abstrait associé à un concept est jugé correct par des experts du domaine et que les noms de relations proposés pour désambiguïser la relation « est lié à » sont également considérés corrects.

Mots clefs : Thesaurus, Ontologie, Création de ressources, Exploration de textes.

Abstract

Abstract text. 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.

Our proposals rest on the one hand on knowledge present in the thesaurus and on that which we automatically extract from a corpus of reference. Thus, certain relations between terms present in the thesaurus are directly exploited to formalize knowledge in the form of ontology (relations "To use rather", "To use to indicate", "is generic than"). Other knowledge is directly extracted from the analysis of the corpus (new labels and hierarchical bonds in particular). Lastly, a generic resource is used to define abstract types making it possible to treat on a hierarchical basis the high level concepts. These various resources are used in a complementary way to finish the enrichment of the ontology (clarification of the relation "is related to" from the thesaurus). The results obtained on a sample of thesaurus in Astronomy show that the method is relevant since 89% of the abstract types associated to concepts are judged as correct by domain experts and most of the proposed labels to disambiguate the "is related to" relationship are also judged as correct.

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 transporter en un langage plus strict (langage documentaire) le langage naturel utilisé dans les documents et dans les requêtes [9]. 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. 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 [34]. 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 [33], leur radicalisation [31], la reformulation de requête automatique par réinjection de pertinence [23] ou par ajout de termes co-occurrents [32] 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 compléter la connaissance extraite des contenus des documents, les systèmes d'indexation automatiques se sont également intéressés à l'utilisation des thésaurus. Gauch et Smith [18] étend automatiquement ainsi les requêtes des utilisateurs en se basant 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 [24] ou le système IRAIA [12] 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, d'autre part 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 [30]. Différentes solutions pour palier à ces inconvénients ont été proposées dans la littérature. La construction automatique de

[1] Un document secondaire est un document issu de la description de documents primaires.
[2] TREC : Text REtrieval Conference http://trec.nist.gov

thésaurus peut ainsi faire appel à des techniques basées sur le calcul de corrélations entre termes [41], la classification automatique de termes [7], la classification de documents [10] ou des approches prenant en compte des connaissances linguistiques [19]. 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 et représentées de façon formelle en se basant le langage OWL[4] et en rendant 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 semble être prometteuse. En effet, les thésaurus 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 [37] qui pourtant joue un rôle primordial dans la communication homme-machine [30]. Les ontologies permettent de reconsidérer ce problème puisqu'il s'agit d'une « *spécification formelle et explicite d'une conceptualisation partagée* » [13]. Cependant, leur élaboration est coûteuse ; elle nécessite de nombreuses interventions manuelles. En effet, les techniques de construction d'ontologies de la littérature ne basent l'élaboration de l'ontologie sur aucune connaissance préalable du domaine mais sur un corpus de référence qui est analysé [21], [1].

Notre approche à 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.

Ce rapport 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 décrit notre méthodologie. Cette section explicite les problématiques auxquelles la méthode doit répondre et ses différentes étapes. La section 5 présente les mécanismes utilisés pour créer le niveau d'abstraction conceptuel à partir du thésaurus. La section 6 explique comment la structure de l'ontologie (liens entre concepts) est construite. Enfin, la section 7 présente l'évaluation de notre approche dans le cadre de l'astronomie.

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

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

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

2. THESAURUS ET ONTOLOGIES

2.1. Normalisation du contenu d'un thésaurus

Les normes ISO 2788 et ANSI Z396 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 [16] : relations d'équivalence, hiérarchiques et associatives. Le schéma conceptuel (formalisme diagramme de classes de UML) du contenu d'un thésaurus est présenté dans la figure 1.

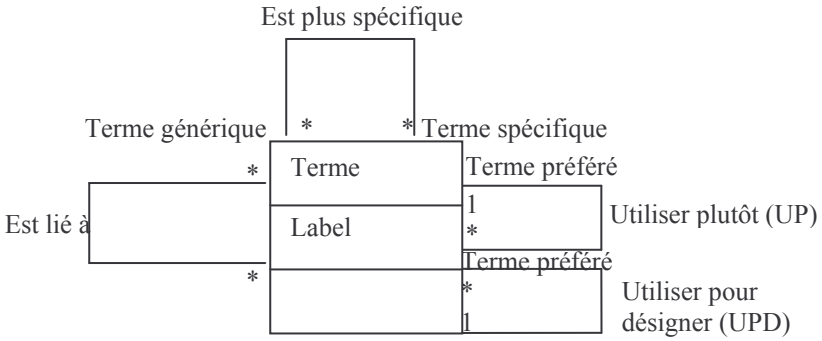


FIG. 1. - *Relations entre termes dans un thésaurus*

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 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 » [15]. La relation associative « est lié à » est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques [40]. Par exemple, dans le thésaurus BIT7 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

[6] http://www.techstreet.com/cgi-bin/pdf/free/228866/z39-19a.pdf
[7] http://www.ilo.org/public/libdoc/ILO-Thesaurus/french/tr1740.htm

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'association entre concepts sont sémantiquement mieux décrits.

Le schéma conceptuel d'une ontologie est présenté dans la figure 2 et est décrit dans ce qui suit.

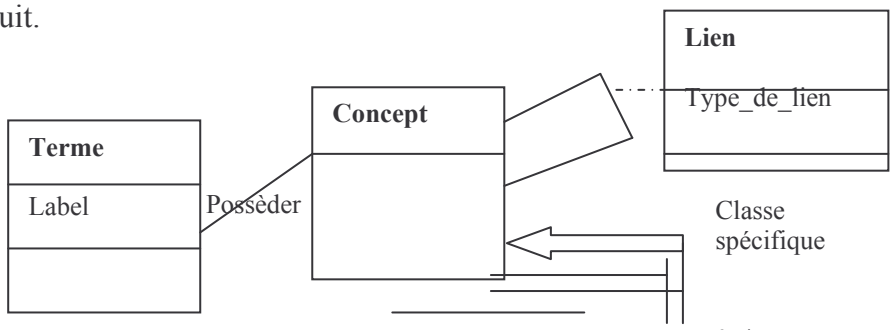


FIG. 2. – *Relations entre concepts dans une ontologie*

2.2.1. Concept et label textuel

Un concept est identifiée à partir d'un identifiant unique. Les labels d'un concept correspondent aux possibles unités lexicales associées à un concept. Dans le schéma conceptuel proposé dans Miles et al. [29], 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 Recherche d'Information (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 une même propriété.

2.2.2. Relation entre concepts

Les concepts sont organisés à partir de relations taxonomiques. Les concepts peuvent aussi être reliés entre eux à partir de relations non taxonomiques. Ce type de relations permet de lier deux concepts en spécifiant le concept de départ de la relation et le concept d'arrivée. Des propriétés peuvent être ajoutées à la relation, telles que la transitivité, la symétrie et la fonctionnalité ou l'inverse d'une autre relation.

2.2.3. Formalisation

La structure d'une ontologie légère est un tuple $S := \{C, R, A, T, CAR_R, \leq^C, \sigma_R, \sigma_A, \sigma_{CAR}\}$ 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,

  $\leq^C (c_1, c_2)$ signifie que c_1 subsume c_2 (relation orientée)
- $\sigma_R : R \rightarrow C \times C$ est la signature d'une relation associative,
- $\sigma_A : A \rightarrow C \times T$ est la signature d'une relation d'attribut,

- σ_{CARR} : R→ CAR_R spécifie la caractéristique d'une relation association.,

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,

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

Le langage OWL 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.

2.2.4. Intérêt des ontologies pour la recherche d'information

Dans la littérature, de nombreux travaux récents portent sur l'utilisation d'ontologies pour l'aide à la recherche d'information. Ils s'appuient sur des ressources généralistes comme WordNet ou des ressources plus spécialisées. Buscaldi et al [8] proposent d'étendre les requêtes des utilisateurs en s'appuyant sur la ressource WordNet et plus spécifiquement sur les relations de synonymie et de méronymie. De la même façon, Baziz et al. [3] propose une méthode d'expansion de requête en utilisant cette même ressource, mais en y ajoutant un processus de désambiguïsation afin que l'expansion n'amène pas trop de bruit. Guo et al. [22] et Hersh et al. [24] basent l'expansion sur la ressource UMLS pour la recherche dans la collection Medline. Ces méthodes se basent sur le fait que l'expansion à partir de connaissances explicites entre les termes permet d'étendre la requête de façon plus sémantique qu'une approche basée sur la co-occurrence des mots dans les textes. Ce même type d'hypothèse est à l'origine de l'indexation sémantique. Dans ce type d'indexation, les termes retenus pour représenter un document ne sont plus les termes issus des documents, mais les concepts associés à ces termes. Dans Mihalcea et Moldovan [28], l'indexation est basée à la fois sur les termes des documents, mais également sur les Synset associés dans WordNet. Cette approche implique un processus de désambiguïsation dans la mesure où la ressource est généraliste. L'indexation sémantique peut également s'appuyer sur une ressource du domaine associé à la collection comme dans Hernandez [25].

Que ce soit pour l'indexation ou la reformulation de requête, l'utilisation de concepts plutôt que de termes permet d'aboutir à des systèmes possédant une sémantique plus fine dans la mesure où l'ambiguïté des termes et les relations entre termes sont connues à priori. Cependant, ces mécanismes font l'hypothèse de l'existence de telles ressources.

La disponibilité de ce type de ressources sous format normalisé est donc un enjeu important dans le domaine de la recherche d'information. Cependant, l'élaboration d'une ontologie est coûteuse. En effet, les techniques de construction d'ontologies de la littérature basent généralement l'élaboration de l'ontologie sur aucune connaissance préalable du domaine et s'appuient sur de nombreuses interventions manuelles.

Nos travaux portent donc sur la construction d'ontologies à partir de thésaurus existants, enrichis par une analyse d'un corpus de référence. Ces travaux sur la migration des thésaurus présentent différents avantages. L'uniformisation de la représentation des ressources grâce à des langages dédiés au web sémantique (tels que RDF et OWL) permettra de distribuer facilement ces ressources sur le web, leur manipulation en sera facilitée mais surtout cela permet d'envisager un enrichissement sémantique incrémental. Les outils dédiés aux ontologies permettront de les visualiser, de les annoter de façon uniforme. A partir de telles ressources, les applications de recherche d'information pourront s'appuyer sur ces ressources élémentaires simples, sans avoir à faire face à l'hétérogénéité des formats.

3. METHODES DE CONSTRUCTION D'ONTOLOGIES : ETAT 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 [11]. 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 [27], différents types d'approches sont distingués en fonction du support sur lequel elles se basent : à partir de textes, de dictionnaires, 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 basées sur les textes dans la mesure où elles ouvrent des perspectives, selon nous, à un enrichissement incrémental de ressources.

Concernant la conception d'ontologies à partir de textes, différents outils de conception ont été développés8. 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 [27]. Text-To-Onto est intégrée à la plate-forme logicielle KAON qui permet l'édition et la maintenance d'ontologies [6]. 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 [17]. 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. Dans le projet TOVE de Gruninger et Fox [20], l'ontologie de domaine est construite (manuellement) à partir des scénarios d'entreprises pour lesquels elle sera utilisée. Cette méthodologie reste sommaire et aucune étape n'est décrite par rapport aux techniques qui peuvent y être employées. De plus, elle est spécialisée sur la spécification d'ontologies pour les entreprises. En revanche, les méthodologies METHONTOLOGY de Fernandez et al. [14] et KACTUS (modelling Knowledge About Complex Technical systems for multiple USe) de Schreiber et al. [35] sont conçues pour être appliquées dans des cadres plus généraux. Dans KACTUS, la méthodologie vise à réutiliser des ontologies existantes et propose des mécanismes permettant cette réutilisation. Ce principe

[8] Pour une étude comparative détaillée des outils de conception d'ontologies, nous renvoyons le lecteur aux travaux de Su et Ilebrekke [38] et de Sure et Corcho [39].

est intéressant dans la mesure où il évite de construire une ontologie à partir de rien. METHONTOLOGY s'applique à clarifier les différentes étapes de la construction en respectant des activités de gestion de projets (planification, assurance qualité), de développement (spécification, conceptualisation, formalisation, implémentation, maintenance) et des activités de support (intégration, évaluation, documentation).

La méthodologie TERMINAE de Aussenac et al. [1] 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 RFIEC9. Elle repose sur l'utilisation d'outils de traitement automatique des langues analysant les termes de textes et les relations lexicales. Les termes sont regroupés suivant leur contexte et facilitent la création de concepts et de relations sémantiques. Les concepts et relations sont ensuite formalisés dans un modèle.

Cette méthodologie est composée de plusieurs étapes :

- La première consiste en la description des besoins (utilisation de l'ontologie, connaissance à représenter…),
- L'étape suivante conduit à construire un corpus sur lequel les outils de traitement automatique de langues se reposent,
- La troisième étape correspond à l'étude linguistique. Des outils sont appliqués au corpus afin d'extraire les termes et leurs relations lexicales et syntaxiques. Une application de la méthodologie est proposée par les auteurs à partir des outils LEXTER de Bourigault [4] et Caméléon de Seguela et Aussenac [36]. Le premier extrait les termes candidats à partir de leurs dépendances syntaxiques. Le second extrait des relations entre termes à partir de patrons linguistiques,
- La phase suivante, appelée phase de normalisation, vise à conceptualiser les résultats de l'étape précédente. Les termes à conserver sont sélectionnés en fonction de leur contexte et définis à partir d'une définition en langage naturel. Les concepts sont ensuite identifiés ainsi que les relations sémantiques entre eux. Ils sont représentés sous forme d'un réseau sémantique,
- La dernière étape est celle de la formalisation. Le réseau sémantique précédemment obtenu est traduit et enrichi dans un langage formel.

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.

[9] 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)

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 les contenus des granules dans la phase d'indexation sémantique et le contenu 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 seront 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 reposera 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 reposera 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 reposera 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. Le corpus doit décrire les éléments de connaissance qui seront intégrés automatiquement dans l'ontologie. Dans le cas de la transformation basée sur un thésaurus, le corpus est un complément qui 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. Il est évident que des connaissances pourront être rajoutées de façon manuelle à l'ontologie, mais dans ce rapport nous ne nous intéressons qu'à des connaissances extraites automatiquement des ressources (thesaurus et textes). Dans le cadre auquel nous nous intéressons, l'ontologie est créée pour des activités de RI ; le corpus considéré doit aider à préciser le contexte des documents du domaine d'intérêt considéré. 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 linguistique du corpus. Cette étape vise à extraire à partir 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 et les relations entre termes qui ne sont pas explicitées dans le

thésaurus. Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique SYNTEX10 [15]. 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.

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

4. CONCEPTUALISATION DU LEXIQUE DU THESAURUS

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.

4.1. Regroupement des termes en concepts

4.1.1. 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 de chacun 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)**

4.1.2. 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.

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 3 schématise plusieurs exemples de groupements. Pour faciliter la lisibilité, les termes préférés sont en gras majuscules.

Extrait d'un thésaurus (IAU)

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

Exemple de termes regroupés par R2
zenith tubes UP **ZENITH TELESCOPES**
PHOTOGRAPHIC ZENITH TUBES UP zenith tubes

FIG.3 - *Exemples de groupements des termes du thésaurus*

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

4.1.3. 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)

4.2. 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 (le terme sera inséré au pluriel) et les termes désignant des objets indénombrables (le terme sera inséré au singulier). Ces règles sont scrupuleusement respectées dans la plupart des thésaurus, comme dans le thésaurus de l'astronomie IAU.

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. Notre choix s'est plutôt porté sur l'utilisation de la ressource lexicale WordNet. La figure 4 illustre les concepts identifiés dans la figure 3 pour lesquels les labels sont mis au singulier.

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

FIG. 4 - *Exemples de concepts labellisés par des termes au singulier*

5. 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.

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 5 schématise ces différents mécanismes.

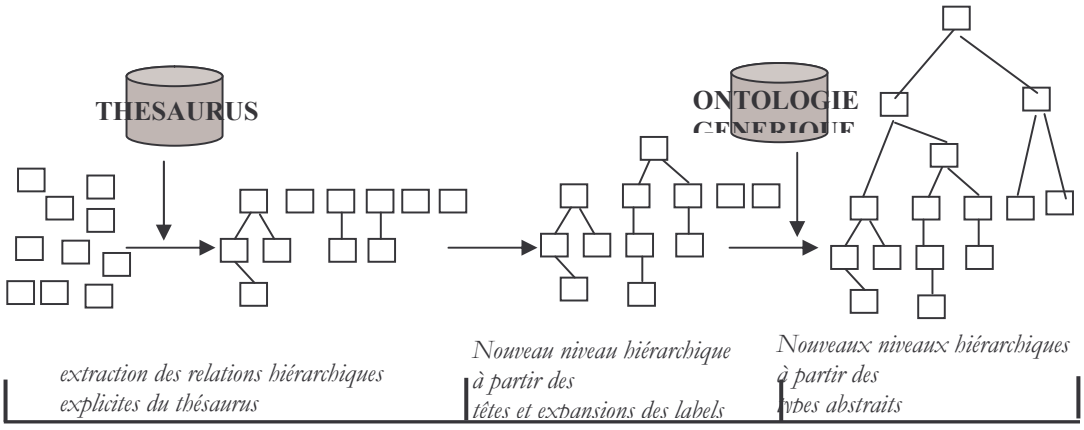


FIG. 5 - *Mécanisme de construction de la hiérarchie de concepts*

5.1.1. 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 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. 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)

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 : 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 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 planetary nebula et nebula est donc inutile.

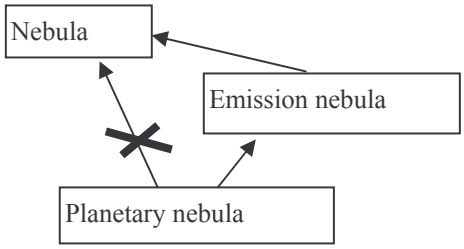


FIG. 6 - *Exemple de redondance dans la hiérarchie de l'ontologie*

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 ≠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 [37]. 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. 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. 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.

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 Soergel et al. [37], 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.

5.1.3.1. 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 également suivie dans OntoLearn de Velardi [42] 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 7). 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 8). 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)

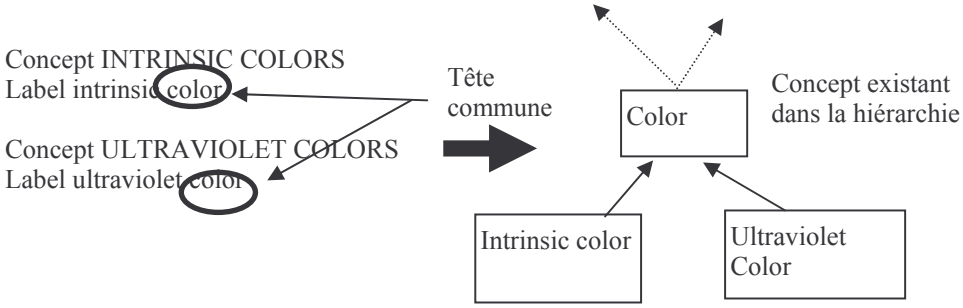


FIG. 7. - *Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie*

Si tete(F^{-1}(c_1))= tete(F^{-1} (c_2)) alors si tete(F^{-1} (c_1))$\notin$ 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)

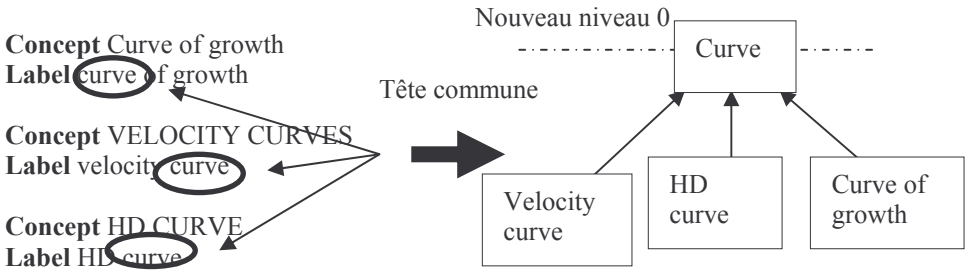


FIG. 8.- *Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie*

5.1.3.2. 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*

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. L'idée est donc de s'appuyer sur la connaissance modélisée dans cette ontologie générale pour créer un niveau hiérarchique supérieur dans l'ontologie générée automatiquement.

Pour cela,, les concepts le plus généraux de la nouvelle ontologie doivent être 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. 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 retenu (relation hyperonymie dans WordNet),
- les Synsets ancêtres du Synset retenu dans WordNet (relation hyponymie dans WordNet),
- les labels des concepts descendants du concept dans l'ontologie (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 9 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. 9.- *Extrait du nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie*

6. 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.

6.1. 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)

6.1.1. 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.

6.1.2. 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. 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
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*

6.1.3. 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 et al. [37]. 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.

6.2. 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 [5]. 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 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 $c_2 \in C_{Onto}$
$\Rightarrow G^{-1}(r') \in G^{-1}(r)$

(R10)

7. MISE A JOUR

Les documents du corpus de référence sont utilisés pour mettre à jour la connaissance de l'ontologie. Cette mise à jour a pour but d'ajouter à l'ontologie de nouveaux termes qui n'étaient pas présents dans le domaine lors de la conception du thésaurus et de situer ces termes dans l'ontologie.

7.1. Détection de nouveaux termes

Afin de déceler de nouveaux termes du domaine non présents dans l'ontologie, des termes du corpus sont extraits à l'aide de SYNTEX. Deux pondérations complémentaires permettent de sélectionner les termes à ajouter parmi tous ceux extraits (les termes possédant un poids suffisant par rapport à une de ces deux pondérations sont sélectionnés).

La première pondération est la fréquence totale d'un terme. Elle représente le nombre total d'apparitions du terme dans le corpus. Elle permet d'extraire les termes fréquemment utilisés et donc généraux du corpus (règle R11). La formule utilisée est la suivante :

$$globalité(terme,corpus) = tf_{terme,corpus} \quad \textbf{(1)}$$

où $tf_{terme,corpus}$ représente la fréquence d'apparition d'un terme du corpus

Si $t \in L_{corpus}$ et globalité(t)>seuil => $t \in L_{COnto}$ **(R11)**

La figure 10 présente un échantillon des syntagmes nominaux les plus fréquents du corpus dans le domaine de l'astronomie non présents dans l'ontologie. Ils ont été validés comme manquants (cf section 8).

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

FIG. 10 *Termes généraux de l'astronomie non présents dans le thésaurus*

La deuxième pondération vise à extraire les termes spécifiques du corpus (règle R12). Elle repose sur la mesure tf.idf qui extrait les termes discriminants d'un document. Cette mesure favorise les termes apparaissant dans le document et n'apparaissant pas dans le reste de la collection. Afin de l'appliquer à l'extraction de termes discriminants d'un corpus, la mesure proposée repose sur la moyenne de tf.idf obtenue par les termes sur l'ensemble des documents du corpus.

$$spécificité(terme,corpus) = \underset{\{granule_i\} \in corpus}{moyenne} (tf_{terme,granule_i} \times idf_{terme}) \quad \textbf{(2)}$$

$$idf_{terme} = \log\left(\frac{N}{f_{terme}}\right) + 1$$

où $tf_{terme,granule}$ représente la fréquence d'apparition d'un terme du lexique d'un corpus L_{corpus} dans un granule du corpus
et f_{terme} correspond au nombre de granules contenant ce terme

La figure 11 présente un échantillon des syntagmes nominaux les plus discriminants du corpus retenus pour le domaine de l'astronomie non présents dans l'ontologie et qui ont été validés comme manquants.

Si $t \in L_{corpus}$ et spécificité(t)>seuil
=> $t \in L_{COnto}$
(R12)

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

FIG 11 *Termes spécifiques de l'astronomie non présents dans le thésaurus*

7.2. Intégration des termes dans l'ontologie

Les nouveaux termes détectés par l'étape précédente doivent être intégrés à l'ontologie. Nous nous basons sur le rapprochement des mots composant le nouveau terme avec les labels des concepts de l'ontologie contenant ces mots. Plus spécifiquement, deux procédés sont mis en place.

Le premier consiste à analyser la tête et l'expansion du syntagme retrouvé. La tête et l'expansion sont ensuite recherchées dans les labels des concepts de l'ontologie. Si ces syntagmes appartiennent au lexique de l'ontologie, les concepts correspondants ainsi que leurs types sont extraits.

- Dans le cas où la tête et l'expansion correspondent à des labels, le nouveau terme permet de proposer une nouvelle relation entre ces concepts. La relation proposée dépend du type des concepts et de la matrice de relations entre types abstraits (cf règle R13).
- Dans le cas où seule la tête est retrouvée, le nouveau syntagme est proposé pour être une nouvelle classe fille du concept représenté par la tête. La queue du syntagme permet, dans ce cas là, de spécifier le concept représenté par la tête (cf règle R14).
- Dans le cas où seule l'expansion du syntagme est label de l'ontologie, la tête est proposée pour être label d'un nouveau concept. Le concept générique relatif à la tête est demandé à un expert et la relation entre les deux concepts est établie à partir de la matrice.

Ces règles sont formalisées comme suit :

Soient $t \in L_{corpus}$ à ajouter dans L_{COnto} , $ta_1 \in C_{Onto}$ et $ta_2 \in C_{Onto}$
Si tete(t) $\in L_{COnto}$ avec tete(t) « sous-classe de » ta_1 et queue(t) $\in L_{Conto}$ « sous-classe de » ta_2
Soient $r \in R_{Onto}$ avec σ_{ROnto}: $R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine
=> $r' \in R_{Onto}$ avec $G^{-1}(r') \in G^{-1}(r)$
(R13)

Soient $t \in L_{corpus}$ à ajouter dans L_{COnto} ,
Si tete(t) $\in L_{COnto}$ et queue(t) $\in L_{COnto}$
=> $t \in L_{COnto}$ avec F(t)=c et c « sous-classe de » F(tete(t))
(R14)

Lorsque le premier procédé ne permet pas de proposer de nouvelles relations ou concepts, un nouveau procédé est mis en place. Il consiste à exploiter le contexte d'apparition du syntagme dans le corpus. Les labels de l'ontologie sont recherchés dans le contexte du syntagme. Lorsqu'un label est retrouvé, le syntagme est proposé pour être label d'un nouveau concept. Ce concept est typé par l'expert puis une relation entre les deux syntagmes est proposée à partir de la matrice.

8. 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 Astronomie11. 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 ce rapport). 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.

8.1. 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.

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

8.2. Types abstraits

Au niveau le plus générique de l'ontologie créée après application des règle R1 à R9 de la transformation, 1132 concepts ont été définis. Les différentes étapes de réduction de ce nombre par recherche de concepts généralisant nous conduit après application des règles R6, R7, R8 et R9 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**	**Pourcentage 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.

8.3. Spécification des relations associatives entre concepts

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

8.3.1. 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. Cette spécification 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.

8.3.2. 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. 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 property	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.

8.3.3. 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.

8.4. Pertinence des mises à jour

La méthode que nous proposons permet également de mettre à jour l'ontologie extraite à partir du thésaurus. Cette mise à jour repose sur l'extraction des corpus de référence de nouveaux termes et sur leur intégration à l'ontologie.

8.4.1. Termes ajoutés

Les règles R11 et R12 permettent de détecter de nouveaux termes à ajouter à l'ontologie. La règle R11 extrait les termes généraux non intégrés à l'ontologie. La méthode a été appliquée sur les noms (composés d'un seul mot), mais donne de très mauvais résultats car les termes ainsi extraits se rapportent au vocabulaire consacré à la rédaction de publication. Des exemples de ces mots sont article, author, publication, result … Nous avons analysé la pertinence de la sélection de tels termes à partir des syntagmes nominaux extraits par l'analyseur syntaxique. Les résultats sont distingués en fonction des corpus desquels ils sont extraits.

Sur le corpus publié en 1995, 72% des termes extraits par la mesure de généralité proposée ont été acceptés pour être ajoutés à l'ontologie par les astronomes (le seuil étant fixé pour englober des fréquences allant de la fréquence maximale à 70% au moins). Ceci montre que, bien que le thésaurus soit une ressource terminologique, lors de sa création, certains des termes n'ont pas été capturés. La mise à jour des termes de l'ontologie est donc indispensable. Des expérimentations devront être réalisées pour fixer le seuil optimal.

Sur le corpus publié en 2002, parmi les 100 syntagmes nominaux les plus généraux, 62 sont validés pour être intégrés à l'ontologie. Ces termes soit n'apparaissent pas dans le corpus de 1995, soit apparaissent avec un score de généralité beaucoup plus bas. L'utilisation d'un corpus récent du domaine est donc primordiale pour mettre à jour l'ontologie à partir de termes présents dans des documents publiés dans la même période que les documents à indexer.

La règle R12 extrait des termes spécifiques à la collection. Elle a été testée pour l'extraction de syntagmes nominaux. Sur les 60 syntagmes ayant les plus forts taux de spécificité, 14 sont validés. Les résultats mettent en évidence le fait que les termes spécifiques à la collection doivent être extraits, les astronomes insistent en effet sur la forte pertinence de ces termes. Cependant, la mesure devrait être affinée pour ne pas sélectionner les termes non pertinents.

8.4.2. Proposition de leur placement dans l'ontologie

Deux méthodes correspondant aux règles R13 et R14 ont été proposées pour intégrer à l'ontologie les nouveaux termes détectés. La première vise à intégrer ces termes comme des labels de nouveaux concepts sous-concepts des concepts existants. La seconde permet de créer de nouvelles relations entre les concepts existants. Ces deux approches ont été évaluées sur 10% des nouveaux termes choisis aléatoirement parmi les termes extraits par la mesure de généralité du corpus publié en 1995. Ces termes ont été jugés pertinents par les astronomes. Les résultats de la détection de nouveaux concepts intégrés à l'ontologie en tant que concepts sous-concepts de concepts existants sont présentés dans le tableau 5. Les résultats de l'intégration des nouveaux termes en tant que relations associatives entre concepts existants sont présentés dans le tableau 6.

Pourcentage de nouveaux concepts créés pertinents	Pourcentage de concepts correctement rattachés à l'ontologie
100%	100%

TAB.5 - *Résultat de l'intégration des nouveaux termes dans l'ontologie*

Pourcentage de nouvelles relations entre concepts existants proposées pertinentes	Pourcentage de relations correctement labellisées
68%	62%

TAB 6 *Résultat de l'intégration des nouveaux termes dans l'ontologie*

Les résultats obtenus montrent l'intérêt de nos deux approches. Les nouveaux concepts créés à partir des nouveaux termes sont pour la totalité, pertinents. Les nouvelles relations ainsi que leur label sont pour la plupart correctes. Notons cependant que 30% des nouveaux termes examinés ne sont pas traités par ces deux approches et qu'une méthode devra être proposée pour détecter de nouveaux termes qui pourront être définis comme nouveaux labels de concepts existants.

Cette phase de mise à jour peut impliquer des restructurations de l'ontologie. L'ajout de concepts et de relations peut en effet modifier le sens de certains éléments de l'ontologie. Afin de limiter ce cas de figure, deux considérations sont prises en compte. Lorsqu'un nouveau concept est proposé pour être sous-concept d'un concept existant, les concepts futurs ancêtres et leurs relations définis dans l'ontologie sont présentés à l'expert. Celui-ci ne valide l'ajout d'un nouveau concept que lorsque les liens avec les différents ancêtres et les différentes relations sont corrects. Dans le cas de l'ajout de nouvelles relations, seules sont validées les relations considérées comme essentielles pour chacune des instances du concept. Cette considération se rapproche de la notion de rigidité définie comme méta-propriété pour l'élaboration d'ontologie formelle dans [Guarino 2002]. L'ensemble des méta-propriétés (unité, identité, dépendance) devrait être pris en compte pour vérifier la cohérence de l'ontologie dans le cas où celle-ci nécessiterait un niveau formel de représentation.

9. CONCLUSION

Dans ce rapport, 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. [37] et Wielinga et al. [43] 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 termes, de relations entre concepts). 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).

Une contribution importante présentée dans ce rapport 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. Une méthode devra être proposée 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). Une méthode devra également permettre d'aider l'expert dans le choix des relations entre concepts proposées si celles-ci peuvent avoir plusieurs labels.

10. REMERCIEMENTS

Les travaux présentés dans ce rapport 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 les astronomes du CDS qui ont évalué nos propositions.

11. REFERENCES

[1] 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.

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

[3] M. Baziz, M. Boughanem et N. Aussenac-Gilles. Evaluating a Conceptual Indexing Method by Utilizing WordNet. WorkShop Clef 2005, .

[4] D. Bourigault. Lexter, a Natural Language Processing Tool for Terminology Extraction. EURALEX International Congress, 1996.

[5] 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.

[6] E. Bozsak et al. KAON - Towards a large scale Semantic Web. In Proceedings of the Third International Conference on E-Commerce and Web Technologies (ECWeb), vol. 2455, pages 304–313. Springer-Verlag LNCS, 2002.

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

[8] D. Buscaldi, P. Rosso, M. Montes-y-Gómez: Context Expansion with Global Keywords for a Conceptual Density-Based WSD. CICLing, pages 263-266, 2005.

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

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

[11] [Ding 2002].

[12] 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.

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

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

[15] 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.

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

[17] [Gal 2004].

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

[19] 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.

[20] M. Gruninger, M. Fox, The logic of enterprise modelling. In Brown, J. & O'Sullivan, D., (Eds.), Reengineering the Enterprise, pages 83–98. Chapman and Hall, 1995.

[21] N. Guarino (ed.), Formal ontology in information systems, IOS press, Amsterdam (NL), 1998.

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

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

[24] 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.

[25] N. Hernandez, J. Mothe, An approach to evaluate existing ontologies for indexing a document corpus, Actes de AIMSA, pages 11-21, 2004.

[26] W. Hersh, S. Price, and L. Donohoe, Assessing thesaurus-based query expansion using the UMLS metathesaurus, Journal of Americian Medical Informatics Association, vol. Suppl. S 2000.

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

[28] R.Mihalcea, D.I. Moldovan, Semantic Indexing using WordNet Senses, Actes de ACL Workshop on IR & NLP, acl.ldc.upenn.edu/W/W00/W00-1104.pdf, 2000.

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

[30] 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.

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

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

[33] 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.

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

[35] G. Schreiber, B. Wielinga, W. Jansweijer, The kactus view of the 'o' word. IJCAI'1995, Workshop on Basic Ontological Issues in Knowledge, 1995.

[36] P. Séguéla et N. Aussenac-Gilles, Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, In Actes de la conférence Ingénierie des Connaissances, pages 79-88, 1999.

[37] 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

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

[39] 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.

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

[41] 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.

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