

Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents

Nathalie Hernandez *, †, Claude Chrisment*, Gilles Hubert *,
Josiane Mothe *, †

* IRIT, 118 route de Narbonne, 31062 Toulouse-Cedex 4, France
{hernandez,chrisment, hubert, mothe}@irit.fr

† GRIMM, 5, allées A. Machado, 31058 Toulouse Cedex, France

‡ ERT34 - IUFM, 56 av. de l'URSS, 31078 Toulouse Cedex 4,
France

Résumé

Les techniques d'indexation pour la recherche d'information s'appuient sur l'extraction de termes dans les documents, termes qui servent ensuite de base pour l'accès à ces documents. Ces techniques font l'hypothèse que les documents contiennent toute l'information utile pour leur indexation et ne prennent pas en compte le fait que les termes peuvent être ambigus ni qu'ils possèdent des liens entre eux. L'indexation sémantique (basée sur des concepts au lieu des termes) à partir d'ontologies ne pose pas ces problèmes. Cependant, elle nécessite que les ontologies reflètent la connaissance actuelle du domaine et soient donc régulièrement mises à jour pour une indexation efficace. Nous proposons dans cet article une méthode visant à mettre à jour une ontologie légère de domaine à partir de l'analyse d'un corpus et de la gestion de types abstraits. L'ajout de labels de concept et de liens typés entre concepts est ainsi rendu possible avec un minimum d'intervention humaine. Cette méthode a été évaluée dans le cadre de l'astronomie.

Mots clefs : *Ontologie, Création de ressources, Mise à jour d'ontologies, Langage d'indexation, Exploration de textes.*

Abstract

Indexing techniques in information retrieval are based on the extraction of terms in the documents, terms which are then used to access to these documents. These techniques make the assumption that the documents contain all useful information for their indexing and do not take into account the fact that the terms can be ambiguous nor that they have semantic relationships between them. Semantic indexing (based on the concepts instead of the terms) based on ontologies does not pose these problems. However, it requires that ontologies reflect the current knowledge of the field and thus are regularly updated for an effective indexing. We propose in this article a method aiming at updating a lightweight domain-ontology in a field starting from the analysis of a corpus and the definition of abstract types. The addition of labels of concepts and relationships between concepts is thus made possible with a minimum of human intervention. This method was evaluated within the framework of astronomy.

Key words: Ontology, Creation of resources, Update of ontologies, Indexing language, Exploration of texts.

1. INTRODUCTION

Les systèmes de recherche d'information s'appuient sur une représentation réduite des contenus des documents à base de descripteurs ; ces descripteurs sont ensuite comparés aux termes de la requête d'un utilisateur pour restituer les documents supposés pertinents. Traditionnellement, lorsque le choix des descripteurs est réalisé manuellement, ceux-ci sont issus d'un langage contrôlé ou d'un thésaurus. Un thésaurus est fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance. Dans un thésaurus, 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 : synonymie, relations hiérarchiques et relations non taxonomiques. Ce langage documentaire est ainsi utilisé pour indexer les documents de façon plus homogène. Lorsque l'indexation est au contraire réalisée de façon automatique, les descripteurs sont directement issus des contenus des documents selon une approche qualifiée de « sac de mots » [Salton et McGill, 1983]. Les textes sont analysés : les termes (mots, radicaux, groupes de mots) sont extraits et leur représentativité (ou leur pouvoir

discriminant) est reflétée par un poids qui leur est automatiquement associé [Robertson et Sparck-Jones, 1976]. Contrairement à l'indexation manuelle qui repose sur l'expertise du documentaliste, l'indexation automatique fait l'hypothèse que les documents contiennent toute la connaissance nécessaire à leur indexation. Certaines approches visent à atténuer la frontière existant entre ces deux types d'approche. Dans le cadre de systèmes s'appuyant sur une indexation automatique, plusieurs travaux [Gauch et Smith, 1991] [Jarvelin, 1996] [Mandala, 1999] proposent d'étendre automatiquement les requêtes des utilisateurs en se basant sur des relations entre termes issues d'un thésaurus. Ils s'assurent ainsi d'une meilleure mise en correspondance des requêtes et des représentations des documents (index extraits). Le système Cat a Cone [Hearst et Karadi, 1997] ou le système IRAIA [Englmeier et Mothe, 2003] indexent les documents automatiquement à partir de termes issus d'un thésaurus et s'appuient sur la structure hiérarchique de celui-ci pour permettre à l'utilisateur de naviguer au sein de sa structure et ainsi accéder aux documents associés aux termes. L'utilisation des thésaurus dans des systèmes automatiques pose pourtant un problème crucial dans la mesure où ils ont été conçus pour une utilisation manuelle. Ainsi, certaines connaissances ne sont pas encodées car elles sont supposées connues par leurs utilisateurs (mais pas par les machines). Les thésaurus possèdent un faible degré de formalisation et s'appuient sur la notion de termes plutôt que celle de concepts [Mizoguchi, 2004] ils n'ont pas de niveau d'abstraction conceptuelle qui pourtant joue un rôle primordial dans la communication homme-machine [Soergel et al., 2004]. Enfin, leur format n'est pas normalisé : fichiers ascii, html, bases de données co-existent, ce qui peut poser des problèmes lors de leur utilisation.

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. Les normes en cours d'élaboration dans le cadre du W3C comme SKOS Core¹ visant

¹ <http://www.w3.org/TR/swbp-skos-core-guide/>

à faire migrer les thésaurus vers des ressources plus homogènes et représentées de façon formelle en se fondant sur le langage OWL² ouvrent la voie vers une utilisation plus pertinente des ressources que sont les thésaurus. Ces normes rendent ces ressources disponibles sur le web sémantique. La prise en compte des avancées en ingénierie des connaissances, en particulier au travers des ontologies est également prometteuse. Plusieurs principes [Uschold et Gruninger, 1996] [Guarino, 1998a] et méthodologies [Guarino et C. Welty, 2002][Kassel 2002] 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 production est très coûteux en temps et pose surtout des problèmes de maintenance et de mise à jour [Ding, 2002]. Depuis une dizaine d'années, la conception semi-automatique d'ontologies émerge 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 production d'ontologies. Ces mécanismes mènent généralement à la conception d'ontologies dites légères mais n'incluent pas nécessairement de procédure de mise à jour incrémentale de l'ontologie.

Dans cet article nous proposons une méthode permettant d'aboutir à ce type de processus incrémental. Ainsi, nous nous intéressons aux différents aspects de la mise à jour : ajout de nouveaux termes, de nouveaux concepts et leur lien dans la structure existante (hiérarchie et autres liens sémantiques entre les concepts déjà existants). Ces mises à jour s'appuient sur l'analyse de documents issus du domaine de connaissance modélisé. Ces documents sont analysés via un analyseur syntaxique afin d'extraire les labels de concepts utiles pour l'indexation des documents, mais également les liens potentiels entre ces concepts. Du point de vue de la RI, la mise à jour des ontologies est capitale. Les ontologies utilisées pour indexer sémantiquement les documents doivent permettre aux systèmes d'interpréter leur contenu et donc être en adéquation avec la connaissance abordée dans les documents [Hernandez et Mothe, 2004]. Une méthode de mise à jour incrémentale permet donc de faire évoluer la connaissance représentée dans l'ontologie pour garantir cette adéquation. Nous considérons qu'à des fins de RI, la

² <http://www.w3.org/TR/owl-features/>

suppression de connaissances de l'ontologie n'est pas nécessaire car le système ne prendra simplement pas en compte cette connaissance (les éléments n'étant pas dans les textes ils ne seront pas exploités). Nous nous concentrons sur l'ajout de connaissances extraites des documents pour aider à l'exploitation du corpus.

L'article est structuré de la façon suivante : la section 2 présente des méthodes de mise à jour d'ontologies. Dans les sections suivantes nous présentons nos propositions. La section 3 introduit l'utilisation des types abstraits, qui constituent un niveau hiérarchique de l'ontologie, et indique comment ils peuvent être obtenus. La section suivante s'intéresse à la mise à jour de l'ontologie, au niveau des concepts et de leurs labels ainsi qu'au niveau des relations taxonomiques et non taxonomiques (ou associatives) entre les concepts. Dans cette section, les exemples illustratifs sont issus de travaux réalisés dans le domaine de l'astronomie. Les évaluations présentées dans la section 5 relèvent également de ce domaine.

2. MISE A JOUR D'ONTOLOGIES

Différentes techniques de mise à jour de l'ontologie ont été proposées dans la littérature. Ces techniques visent à extraire de nouveaux termes et à les intégrer dans l'ontologie. Elles reposent sur la détection d'indices lexicaux et statistiques liant les nouveaux termes détectés dans le corpus et les labels de concepts définis dans le lexique de l'ontologie. Afin de nous concentrer sur la mise à jour d'ontologies, nous présentons dans cette section les approches de la littérature prenant en compte une ontologie existante. Pour plus de détails sur les méthodes de construction d'ontologies (extraction de termes, concepts, relations) nous renvoyons le lecteur à [Hernandez 2005].

La méthode proposée dans [Faatz et Steinmetz, 2002] consiste à définir une mesure calculant la distance dans les textes entre les labels de deux concepts de l'ontologie liés par une relation. Cette mesure est ensuite utilisée pour extraire les termes du corpus se trouvant à cette distance des labels des concepts de l'ontologie. Ces termes sont ensuite proposés pour être labels de nouveaux concepts liés par la relation aux concepts existants. La mesure proposée est inspirée de la divergence de Kullback qui calcule la dissimilarité entre deux mots. Elle est étendue pour prendre en compte des indices dans les documents témoins de la relation considérée comme par exemple la relation sémantique « est un ». Cette relation est prise en compte dans la mesure par la fréquence de co-occurrence des labels de deux concepts dans une fenêtre de cinq mots autour des termes. L'avantage de cette approche est de permettre l'apprentissage de la distance représentée par n'importe quelle relation de l'ontologie. Cependant, sa mise en pratique pour tout type de relation est difficile car elle implique de maîtriser les indices du corpus qui permettent de déceler la relation.

Dans [Maedche et al., 2002], la méthode proposée vise à détecter de nouveaux termes mais propose uniquement de les intégrer à l'ontologie par la création de concepts liés par des relations taxonomiques. Une matrice de co-occurrence est constituée à partir des labels désignant les concepts et les termes co-occurrent autour d'eux dans une fenêtre de trois mots dans une même phrase. Ces termes sont ensuite hiérarchisés à partir de méthodes hiérarchiques ascendantes et descendantes. La mesure utilisée pour effectuer les rapprochements entre termes est définie pour prendre en compte l'organisation des termes déjà labels de concepts dans l'ontologie. Les nouveaux termes rapprochés sont proposés pour être

ajoutés à l'ontologie comme sous-concepts des concepts à partir des labels avec lesquels ils sont liés. Les expérimentations présentées dans l'article ne permettent pas de mettre en valeur l'intérêt de l'approche.

Un autre type d'approche vise à mettre à jour les ontologies à partir de la méthode des signatures de thématique [Lin et Hovy, 2000]. Habituellement utilisée pour la production de résumés, cette approche consiste à trouver un ensemble de termes relatifs à une thématique et à pondérer le lien entre chaque terme et la thématique. Appliquée à la construction d'ontologies, cette méthode permet de rapprocher les termes des concepts qu'ils référencent. Deux approches ont été proposées pour mettre à jour Wordnet [Miller, 1988]. WordNet est un système de références lexicales dont la conception a été inspirée par les théories de la mémoire linguistique humaine. WordNet est composé d'ensembles de synonymes appelés synsets, où chaque terme est regroupé en classes d'équivalence sémantique. Chaque ensemble de synonymes représente les labels d'un concept particulier. Un terme peut appartenir à plusieurs synsets et à plusieurs catégories lexicales. Les ensembles de synonymes sont associés par des relations sémantiques : généralité/spécificité, antonymie (relation entre ensembles de mots qui, par leur sens, s'opposent). WordNet couvre le domaine de la langue générale en intégrant le sens des mots dans différents domaines. La méthode proposée dans [Agirre et al., 2000] vise à mettre à jour WordNet par rapport à un corpus donné, en supprimant les labels de concepts (appelés synset dans Wordnet) inutiles et en proposant de nouveaux labels, extraits de documents du Web. La première étape consiste à rechercher des documents du Web relatifs aux concepts de WordNet ; les requêtes sont formulées à partir des termes référençant le concept considéré ainsi que ses hyperonymes. Une collection est ainsi créée pour chaque concept. La deuxième étape consiste à calculer pour chaque terme des documents sa fréquence d'apparition dans les différentes collections. Les termes qui ont une statistique différente dans une collection sont retenus pour constituer les labels associés au concept (aussi appelés dans cette approche « signature du concept » par analogie à signature de thématique). Les termes qui se trouvent dans différentes collections permettent de trouver des liens entre les concepts. Dans [Agirre et al., 2000], l'ontologie ainsi mise à jour est testée sur le corpus Semcor pour une tâche de désambiguïsation. Les résultats qu'elle permet d'obtenir sont plus précis que ceux obtenus par l'utilisation de WordNet non modifié. La limite de cette approche est qu'elle détecte de nouveaux liens entre concepts mais ne permet pas de déterminer comment ces liens doivent être interprétés.

Dans [Alfonseca et Manandhar, 2002], la méthode de la signature des thématiques est également utilisée pour mettre à jour WordNet. La même

démarche d'interrogation du Web est réalisée pour extraire la collection d'apprentissage. La différence est que cette approche vise à proposer de nouveaux concepts et leur placement dans l'ontologie à partir des termes co-occurrent autour des labels associés aux concepts dans WordNet. Le principe de la méthode repose sur l'hypothèse de la distribution sémantique « le sens d'un mot est corrélé au contexte dans lequel il apparaît ». Les termes apparaissant fréquemment dans les documents autour des labels issus de WordNet sont retenus pour être des labels de nouveaux concepts. Une signature de thématique est réalisée pour chacun des concepts (ceux de WordNet ainsi que les nouveaux concepts proposés). Un algorithme parcourt ensuite l'ontologie de sa racine vers ses fils en calculant, au niveau de chaque concept, la similarité entre la signature de ce concept et celle du concept à ajouter. Au niveau de chaque nœud, le fils choisi est celui qui a la similarité la plus forte. L'algorithme s'arrête lorsque le score d'un concept est supérieur à celui de ses fils. Le procédé est entièrement automatique. Il a l'avantage d'extraire de nouveaux termes et de les intégrer directement dans l'ontologie par la création de nouveaux concepts définis à partir de plusieurs labels (les termes de la signature des nouveaux concepts). Bien qu'ils soient obtenus automatiquement, les résultats doivent cependant être validés par un expert.

Les méthodes de mise à jour d'une ontologie se sont essentiellement concentrées sur la détection de nouveaux termes à ajouter à l'ontologie et sur l'intégration de ces termes par la création de concepts rattachés à l'ontologie par la relation « est un ». La détection de relations associatives est un élément important dans la production d'ontologies car elles permettent de spécifier le sens des concepts. Notre approche vise à permettre une mise à jour de l'ontologie en intégrant ce type de relations. Elle repose comme les approches précédemment présentées sur l'analyse de documents textuels du domaine. Afin de mettre en place la construction d'ontologies à partir de textes, il est tout d'abord nécessaire de constituer l'ensemble des documents sur lesquels reposera cette élaboration [Condamines, 2005]. Plusieurs cas de figures peuvent amener à élaborer ce corpus. S'il existe des documents dans lesquels la connaissance peut être capturée, les documents pré-existants sont rassemblés. L'enjeu est alors de collecter des documents existants afin de couvrir le domaine d'intérêt. Une solution, comme nous l'avons vu, est d'interroger le Web à partir de requêtes décrivant le domaine qui devra être traité dans l'ontologie. Une alternative est de choisir un corpus existant et de le valider pour servir de corpus de référence. Dans le cas des travaux portant sur la génération d'ontologies pour la RI, le corpus est généralement composé de l'ensemble des documents à indexer comme

par exemple dans [Koo, 2003]. Si un tel ensemble de documents n'existe pas, des documents doivent être créés spécialement à cet effet. Ce cas de figure se présente quand l'ontologie doit capturer de la connaissance tacite sur un domaine comme, par exemple, lorsque l'ontologie traite de la mémoire d'une entreprise. Le savoir-faire des experts du domaine n'est pas explicitement présenté dans des documents. La connaissance des experts est alors capturée à partir de documents textuels relatant des interviews. La construction de ce type de corpus revient à faire passer les connaissances du tacite à l'explicite.

3. ONTOLOGIE LEGERE DE DOMAINE INCLUANT DES TYPES ABSTRAITS

Notre étude s'intéresse aux ontologies légères de domaine. Nous présentons dans les sections suivantes la formalisation que nous utilisons et qui sera reprise dans la description de la méthodologie de mise à jour que nous proposons. Nous introduisons également la notion de type abstrait et de relation entre ces types. Les types abstraits sont utilisés dans notre approche pour aider à l'insertion de nouveaux concepts dans la hiérarchie de l'ontologie existante ainsi que pour inférer de nouvelles relations entre les concepts de l'ontologie.

3.1. Formalisation d'une ontologie légère

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

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. 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.

Le langage OWL³ 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.

3.2. La construction semi-automatique d'ontologies à partir de thésaurus et de textes

La méthodologie de mise à jour d'ontologies que nous présentons dans cet article s'inscrit dans une méthodologie plus globale qui vise à construire de façon aussi automatique que possible une ontologie. Cette construction s'appuie sur l'utilisation de deux types de ressources : thésaurus et corpus de documents. Ce choix a été motivé par l'existence de nombreux thésaurus dans différents domaines qui ont demandé des efforts importants pour être conçus et qui gagnent à être formalisés au travers des ontologies afin d'en assurer la maintenance, la diffusion et l'utilisation par des applications.

Dans cet article, nous ne développerons pas les propositions que nous avons faites pour transformer un thésaurus et aboutir à une ontologie. Le lecteur intéressé pourra se reporter à [Hernandez 2006]. Cependant, il est important de noter que, dans ce cadre, nous avons été amenés à définir la notion de type abstrait qui permet d'ajouter un niveau d'abstraction

³ <http://www.w3.org/TR/owl-features/>

conceptuelle non défini dans les thésaurus (concepts appartenant au plus haut niveau hiérarchique de l'ontologie).

3.3. Les types abstraits

Dans notre approche, les concepts du plus haut niveau de l'ontologie correspondent aux types abstraits. Un type 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., 2004], ces types 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 concepts de l'ontologie existante à l'un des concepts parmi les centaines de concepts du schéma. Dans notre approche, ces types abstraits sont obtenus de façon automatique à partir d'une ontologie pré-existante [Hernandez, 2006] et de l'utilisation de WordNet. Cette démarche se justifie pleinement dans le cas où l'ontologie initiale est construite de façon semi-automatique à partir d'un thésaurus. En effet ces ressources ne possèdent généralement pas de concepts généraux et les « entrées » de plus haut niveau peuvent être très nombreuses. Par exemple, dans le thésaurus IAU qui sert de validation à notre approche, il y a 1132 « entrées » pour lesquelles aucun terme plus général n'a été défini. Cette démarche peut également avoir son intérêt lors d'une construction semi-automatique d'ontologies à partir de textes. Dans ce cas, il est possible que le corpus ne contienne pas les termes généraux du domaine, connus et sous entendus par l'ensemble des auteurs ; pourtant ces termes généraux peuvent avoir leur importance lors d'une recherche ou lorsque la collection à interroger est multi-domaine.

La définition des types abstraits vise à identifier les concepts dont dépendent les concepts du niveau 0 de l'ontologie générée à partir de corpus ou à partir d'un thésaurus. Cette définition comporte deux étapes. Dans un premier temps, il s'agit de définir les types abstraits du domaine, puis de les intégrer aux concepts de l'ontologie via les liens de subsomption et des liens non taxonomiques. La figure schématise le procédé mis en place.

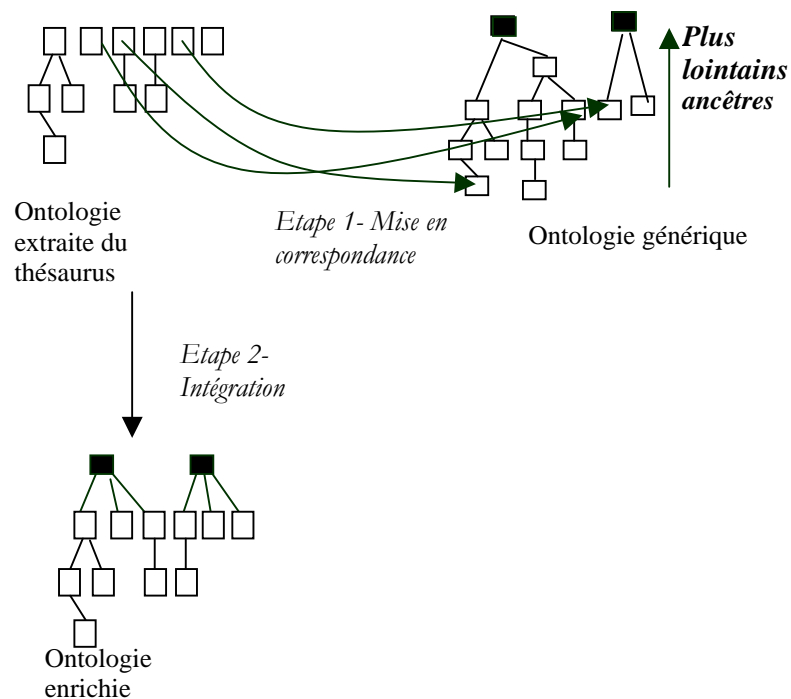


FIG. 1.- Processus d'identification et d'intégration des types abstraits

3.3.1. Identification des types abstraits

Généralement, la définition de types abstraits de haut niveau implique l'intervention humaine. Alternativement, il est possible d'avoir recours 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 ce réseau sémantique indépendant d'un domaine pour créer un niveau hiérarchique supérieur dans l'ontologie initiale (issue dans notre cas de la transformation et de l'enrichissement d'un thésaurus de domaine).

Pour cela, les concepts de plus haut niveau de l'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 généraux associés aux concepts identifiés dans l'ontologie générique. 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 hyperonymie 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 œuvre séquentiellement. Ces trois méthodes permettent d'effectuer la désambiguïsation en fonction du domaine de connaissance décrit dans l'ontologie ainsi que par rapport au contexte local des concepts et des Synsets, que ce soit par rapport à leurs concepts descendants ou ancêtres.

(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 désambiguïé.

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

La figure 2 présente des exemples de types abstraits extraits dans notre cas d'application.

<p>Property : a basic or essential attribute shared by all members of a class</p> <p>Phenomenon : any state or process known through the senses rather than by intuition or reasoning</p> <p>Event : <i>something that happens at a given time</i></p>

FIG. 2.- Extrait du nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

3.3.2. Liens taxonomiques entre types abstraits et concepts de l'ontologie

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 lorsque les labels du concept n'étaient pas dans WordNet, l'association concept/type abstrait est réalisée manuellement.

La règle R1 synthétise les étapes qui conduisent à la définition des types abstraits et à leur rattachement dans la taxonomie induite par l'ontologie.

<p>Si $c \in sw$ avec $sw \in \{\text{Synsets_WordNet}\}$ $\Rightarrow c$ « est sous classe de » ta Avec ta (type abstrait) est le plus spécifique hyperonyme de sw (R1)</p>
--

3.3.3. Liens non taxonomiques 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 d'un corpus de référence.

Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique SYNTEX⁴ [Bourigault et Fabre 2000]. Cet analyseur repose 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).

Ces propositions servent de base à la définition manuelle de relations entre paires de type abstrait et sont synthétisées dans la règle R2.

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 $r'(c_1, c_2)$ 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 »
 $\Rightarrow G^{-1}(r') \in G^{-1}(r)$

(R2)

3.3.3.1. 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 candidats pour caractériser le type abstrait et servir de proposition aux labels des relations associatives entre ces types abstraits. Un expert du domaine valide ensuite l'ensemble des termes référençant la relation. 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 labels de relations ont été validés par l'expert car les instruments astronomiques sont utilisés pour observer ou mesurer les autres concepts du domaine.

⁴ SYNTEX a été développé par D. Bourigault, membre de l'équipe ERSS et partenaire de la plateforme RFIEC.

3.3.3.2. Définition de relations entre types

La définition des relations sémantiques est réalisée entre chaque paire de types abstraits. Une matrice à double entrée est 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 types abstraits deux à deux et reporte les labels qu'il choisit dans les cases de la matrice.

	Property	Phenomenon	Event	Science
Property	<i>Influences</i> <i>Is influenced by</i> <i>Determined by</i> <i>Determines</i> <i>Exclude</i> <i>Has part</i> <i>Is part</i>	<i>Is a property of</i> <i>Induces</i>	<i>Is a property of</i> <i>of</i> <i>induces</i>	<i>Is studied by</i>
Instrumentation	Makes Observes	<i>Observes</i> <i>Measures</i>	<i>Observes</i> <i>Measures</i>	<i>Is Used to studied</i>

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

4. MISE A JOUR DE L'ONTOLOGIE

4.1. Détection de nouveaux labels et concepts

Afin de déceler de nouveaux termes du domaine non présents dans l'ontologie, des termes du corpus sont extraits à l'aide de SYNTAX. Deux pondérations qui sont complémentaires permettent de sélectionner

les termes à ajouter parmi tous ceux qui ont été 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 R3). La formule utilisée est la suivante :

$$\text{globalité}(\text{terme}, \text{corpus}) = \text{tf}_{\text{terme}, \text{corpus}} \quad \text{(1)}$$

où $\text{tf}_{\text{terme}, \text{corpus}}$ représente la fréquence d'apparition d'un terme du corpus

$\text{Si } t \in L_{\text{corpus}} \text{ et } \text{globalité}(t) > \text{seuil} \Rightarrow t \in L_{\text{C}_{\text{Onto}}} \quad \text{(R3)}$
--

La figure 3 présente un échantillon des syntagmes nominaux les plus fréquents extraits automatiquement du corpus dans le domaine de l'astronomie et non présents dans l'ontologie initiale. Ils ont été validés comme manquants par des astronomes.

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

FIG. 3. Termes fréquents de l'astronomie non présents dans l'ontologie issue du thésaurus

La deuxième pondération vise à extraire les termes spécifiques du corpus (règle R4). Elle repose sur la mesure tf.idf qui extrait les termes discriminants d'un document. Cette mesure favorise les termes apparaissant fréquemment dans un document mais dans peu d'autres documents. 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. Des extensions de la mesure tf.idf ont été proposées pour prendre par exemple en compte la taille des documents [Robertson, 1996]. Mais nous souhaitons dans un premier temps évaluer la mesure sous sa forme la plus simple.

$$\text{spécificité}(\text{terme}, \text{corpus}) = \text{moyenne}_{\{\text{document}\} \in \text{corpus}} (t_{\text{terme}, \text{document}} \times id_{\text{terme}}) \quad (2)$$

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

où $t_{\text{terme}, \text{document}}$ représente la fréquence d'apparition d'un terme du lexique d'un corpus L_{corpus} dans un document du corpus et df_{terme} correspond au nombre de document contenant ce terme

La figure 4 présente un échantillon des syntagmes nominaux les plus discriminants du corpus retenu pour le domaine de l'astronomie non présents dans l'ontologie et qui ont été validés comme manquants par les experts (cf section évaluation).

<p style="text-align: center;">Si $t \in L_{\text{corpus}}$ et $\text{spécificité}(t) > \text{seuil}$ $\Rightarrow t \in L_{\text{COnno}}$</p> <p style="text-align: right;">(R4)</p>

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

FIG. 4 Termes spécifiques de l'astronomie non présents dans le thésaurus

4.2. Intégration dans la hiérarchie de l'ontologie

Les nouveaux termes détectés par l'étape précédente doivent être intégrés à l'ontologie. Concernant l'intégration au niveau de la hiérarchie représentée par les liens taxonomiques entre concepts de l'ontologie, deux procédés sont mis en place. Ils reposent sur le rapprochement des mots composant le nouveau terme avec les labels des concepts de l'ontologie contenant ces mots. Plusieurs cas de figure se présentent : il est possible de rattacher le nouveau terme à un concept déjà existant par un lien « is-a » ou bien ce rattachement n'est pas possible car il s'agit par exemple d'un nouveau pan du domaine qui est ajouté via le corpus analysé.

4.2.1. Nouvelle sous classe d'un concept existant

La tête et l'expansion du syntagme nouvellement extrait du corpus sont recherchées dans les labels des concepts de l'ontologie.

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 R5).

Soient $t \in L_{\text{corpus}}$ à ajouter dans L_{COnto} ,
Si $\text{tete}(t) \in L_{\text{COnto}}$ et $\text{queue}(t) \notin L_{\text{COnto}}$
 $\Rightarrow t \in L_{\text{COnto}}$ avec $F(t)=c$ et c « sous-classe de » $F(\text{tete}(t))$
(R5)

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 type abstrait relatif à la tête est demandé à un expert.

4.2.2. Nouveaux concepts et création des concepts pères

Les nouveaux termes à inclure sont regroupés à partir de la tête des termes. 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 5). 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 6). Ce mécanisme permet de créer un nouveau premier niveau de la hiérarchie contenant un nombre plus réduit de concepts.

Si $\text{tete}(F^{-1}(c_1)) = \text{tete}(F^{-1}(c_2))$ alors si $\text{tete}(F^{-1}(c_1)) \in L_{\text{Onto}}$
 $\Rightarrow c_1$ « est une sous classe de » $F(\text{tete}(F^{-1}(c_1)))$
et c_2 « est une sous classe de » $F(\text{tete}(F^{-1}(c_1)))$
(R6)

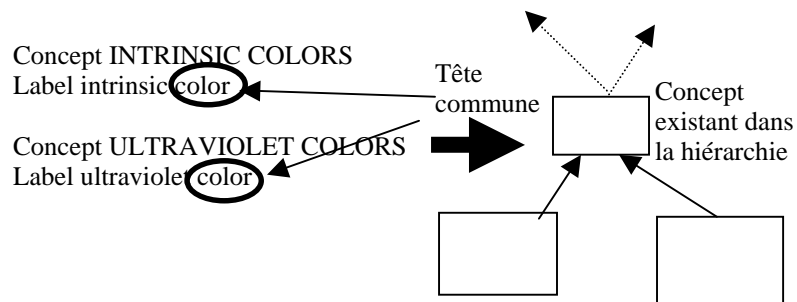


FIG. 5. - Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie

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

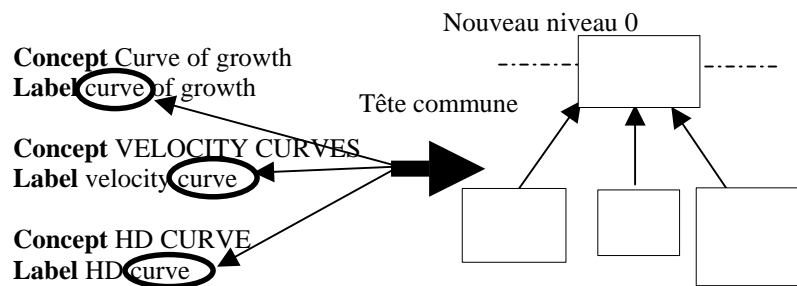


FIG. 6.- Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

4.3. Nouveaux liens non taxonomiques entre concepts

Pour créer des liens entre concepts de l'ontologie, nous nous appuyons sur les relations entre types abstraits.

4.3.1. Liens entre concepts existants

Dans le cas où la tête et l'expansion du terme extrait du corpus correspondent chacun à un label de concepts de l'ontologie, le nouveau terme permet de proposer une nouvelle relation entre ces concepts. La relation proposée dépend d'une part du type des deux concepts référencés par les labels et d'autre part de la matrice de relations entre types abstraits (cf règle R8). Par exemple, l'extraction du terme « observatory telescope » correspondant à deux labels de concepts de l'ontologie (observatory de type abstrait artifact et telescope de type abstrait instrumentation) mène à la proposition d'une relation entre ces deux concepts dont le label est « is a part of ».

Soient $t \in L_{\text{corpus}}$ à ajouter dans L_{COnto} , $ta_1 \in C_{\text{Onto}}$ et $ta_2 \in C_{\text{Onto}}$
Si $tete(t) \in L_{\text{COnto}}$ avec $F(tete(t))$ « sous-classe de » ta_1 et $F(queue(t)) \in L_{\text{COnto}}$ « sous-classe de » ta_2
Soient $r \in R_{\text{Onto}}$ avec $\sigma_{R_{\text{Onto}}}: R_{\text{Onto}} \rightarrow C \times C$ et $r(ta_1, ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine
 $\Rightarrow r' \in R_{\text{Onto}}$ avec $G^{-1}(r') \in G^{-1}(r)$ **(R8)**

4.3.2. Contexte d'apparition dans le corpus

Cette approche consiste à exploiter le contexte d'apparition du syntagme dans le corpus. Sur la base de la matrice de relations entre types abstraits, 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. Nous nous appuyons sur l'analyse distributionnelle réalisée par le module UPERY de SYNTAX [Bourigault, 2002]. Ce type d'analyse consiste à rapprocher des syntagmes en fonction de la ressemblance de leur contexte syntaxique dans le corpus. Les syntagmes obtenus par 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 rapprochés 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 contexte. 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 dans 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 reflétant les relations possibles entre types abstraits (règle R9).

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 « is a characteristic » est proposée entre « *luminosity* » et « *galaxy* » (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 $c_2 \in C_{Onto}$
 Et c_1 « sous-classe de » ta_1 et c_2 « sous-classe de » ta_2
 $\Rightarrow G^{-1}(r') \in G^{-1}(r)$ **(R9)**

5. EVALUATION

Nous présentons dans cette section un retour d'expérience sur l'application de la méthode de transformation en ontologie légère et d'enrichissement d'un thésaurus ; 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 Astronomie⁵. 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 n'est pas le cœur de cet article. Nous nous intéressons ici plus spécifiquement à l'enrichissement de l'ontologie obtenue en se basant sur une analyse de corpus. 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 de 1223 articles publiés en 1995, date de création du thésaurus. Nous l'avons utilisé pour enrichir la connaissance initiale. Le deuxième corpus est constitué de 1834 articles publiés en 2002. Ce corpus a été choisi pour évaluer la mise à jour des connaissances du domaine à partir de documents récents. Les deux corpus ont été validés par des experts du domaine pour décrire les connaissances à représenter dans l'ontologie. Ils contiennent au total 91137 syntagmes nominaux.

5.1. 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.

⁵ <http://cdsweb.u-strasbg.fr/MDA/mda.html>

5.2. Spécification des relations associatives entre concepts

5.2.1. 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 dont 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.

5.2.2. 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 R9 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-occure. 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 2).

	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.2 - *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.

5.3. Nouveaux termes

5.3.1. Termes ajoutés

Les règles R3 et R4 permettent de détecter de nouveaux termes à ajouter à l'ontologie. La règle R3 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 R4 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.

5.3.2. Proposition du placement des nouveaux éléments dans l'ontologie

Deux méthodes correspondant aux règles R8 et R5 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 3. Les résultats de l'intégration des nouveaux termes en tant que relations associatives entre concepts existants sont présentés dans le tableau 4.

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

TAB 3 *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 4 *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. Ce résultat est lié au corpus considéré et nous souhaitons le comparer par l'application de notre méthode à d'autres domaines. Les nouvelles relations ainsi que leurs labels 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 et Welty, 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.

6. CONCLUSION

Dans cet article, nous avons proposé une méthodologie permettant de mettre à jour la représentation sémantique d'un domaine ainsi que sa formalisation. Cette méthodologie vise à extraire, à partir de l'analyse syntaxique de documents du domaine, de nouveaux termes qui sont intégrés à l'ontologie, soit par la création de nouveaux concepts, soit par la création de nouvelles relations entre concepts. Les nouveaux termes sont détectés à partir de deux mesures extrayant les termes spécifiques et les termes généraux non présents dans l'ontologie. L'intégration de ces termes dans la structure de l'ontologie existante 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 aux nouveaux termes détectés dans le corpus. Notre méthode a l'avantage de minimiser le travail de validation par les experts car il repose uniquement sur la validation de propositions.

Cette mise à jour incrémentale peut servir à améliorer un processus d'indexation sémantique lorsque de nouvelles connaissances non représentées dans l'ontologie initiale sont abordées dans la collection à indexer. L'évaluation de la méthode sur le domaine de l'astronomie a montré son intérêt. Elle extrait efficacement des types abstraits qui sont associés aux concepts les plus généraux de l'ontologie. Les nouveaux termes extraits par nos mesures sont globalement pertinents comme le sont les propositions d'intégration des termes dans l'ontologie.

A la suite de cette évaluation, plusieurs perspectives sont envisagées. Nous envisageons de développer une interface de validation de nos propositions de façon à rendre plus agréable le travail des experts. Nous souhaitons également appliquer notre approche à de nouveaux domaines afin de comparer son efficacité sur différents corpus et différentes ontologies existantes. Nous envisageons également 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].

7. 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.

REFERENCES

- E. Agirre, O. Ansa, E. Hovy, D. Martinez, Enriching very large ontologies using the WWW, In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00), 2000.
- E. Alfonseca et S. Manandhar, Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures, EKAW-2002, Lecture Notes in Artificial Intelligence 2473, Springer Verlag, 2002.
- D. Bourigault. Lexter, a Natural Language Processing Tool for Terminology Extraction. EURALEX International Congress, 1996.
- 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.
- D. Bourigault et C Fabre, Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaire, 25, Université Toulouse le Mirail, pages 131-151, 2000.
- A. Condamines, Sémantique et Corpus, Hermès Science Publications, ISBN 2-7462-1055-X, 2005.
- 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.
- A. Faatz et R. Steinmetz, Ontology enrichment with texts from the WWW, In Proceedings of the 2nd Semantic Web Mining Workshop at ECML/PKDD, 2002.
- S. B. Fensel, Knowledge Engineering : Principles and Methods. Data and Knowledge Engineering, 25, pages 161-197, 1998.

- J.-Y. Fortier et 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), pages 39-45, 2004.
- S. Gauch et J.B. Smith, Search Improvement via Automatic Query Reformulation, *ACM Transactions on Information Systems*, 9(3), pages 249-280, 1991.
- N. Guarino, Some ontological principles for designing upper level lexical resources, In Proceedings of the 1st International Conference on Language Resources and Evaluation, 1998.
- N. Guarino et C. Welty, Evaluating Ontological Decisions with OntoClean, In *Communication of the ACM*, 45(2), pages 61-65, 2002.
- 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, pages 29-41, 2001.
- M.A. Hearst et 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.
- N. Hernandez et J. Mothe, An approach to evaluate existing ontologies for indexing a document corpus, *Actes de AIMSA*, pages 11-21, 2004.
- N. Hernandez et J. Mothe, Enrichissement sémantique pour la recherche d'information : méthodologie de transformation d'un thésaurus en une ontologie de domaine, *Revue électronique ARTIST*, Vol 1, à paraître.
- K. Jarvelin, J. Kristensen, T. Niemi, E. Sormunen, H. Keskustalo, Expansion Tool: a deductive data model for thesauri and query expansion, *Finnish information studies FIS-1996-5*, Department of Information Studies, University of Tampere, 1996.
- G. Kassel, OntoSpec : une méthode de spécification semi-informelle d'ontologies, In *Actes des 13èmes journées francophones d'Ingénierie des Connaissances (IC)*, pages 75-87, 2002.
- C.Y. Lin et E.H. Hovy, The Automated Acquisition of Topic Signatures for Text Summarization, In Proceedings of the COLING Conference, 2000.
- A. Maedche, V. Pekar, S. Staab, Ontology learning part one – on discovering taxonomic relations from the web, In *Web Intelligence*, Z. Ning et al (Eds.), Springer, 2002.
- R. Mandala, T. Tokunaga, H. Tanaka, Combining multiple evidence from different types of thesaurus for query expansion, In Proceedings of the 22nd

International ACM SIGIR conference on Research and Development in Information Retrieval, pages 191-197, 1999.

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

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.

S. Koo, S.Y. Lim, S.J. Lee, Building an Ontology based on Hub Words for Informational Retrieval, In Proceedings of the IEEE/WIC International Conference on Web Intelligence, 2003.

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

S. E. Robertson, et S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of SIGIR 1994, pages 232-241, 1994.

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

G. Salton et M. J. McGill. Introduction to Modern Retrieval. McGraw-Hill Book Company, 1983.

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.

M. Uschold, M. Gruninger, Ontologies: principles, methods, and applications, Knowledge Engineering Review, 11(2), pages 93-155, 1996.

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