Laboratoire IRIT – Pôle SIG-EVI

THESE

Présentée devant

L'Université Paul Sabatier de Toulouse

en vue de l'obtention du

Doctorat de l'Université Paul Sabatier
Spécialité Informatique

Par
Nathalie HERNANDEZ

ONTOLOGIES DE DOMAINE POUR LA MODELISATION DU CONTEXTE EN RECHERCHE D'INFORMATION

Soutenue le **mardi 06 décembre** devant le jury composé de :

Claude Chrisment	Professeur à l'Université de Toulouse III (directeur de recherche)
Josiane Mothe	Professeur à l'Institut Universitaire de Formation des Maîtres de Midi-Pyrénées (directeur de recherche)
Gilles Kassel	Professeur à l'Université de Picardie (rapporteur)
Dominique Rieu	Professeur à l'IUT2 de Grenoble (rapporteur)
Ollivier Haemmerlé	Professeur à l'Université de Toulouse II (examinateur)
Fionn Murtagh	Professeur à l'Université de Londres (examinateur)
Nathalie Aussenac-Gilles	Chargé de recherche CNRS (invité)
Françoise Genova	Directrice de l'Observatoire de Strasbourg (invité)

INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, Tél : 05.61.55.66.11

Remerciements

Enfin, cette fameuse page qui me tient tellement à cœur …

Je remercie tout d'abord Monsieur Louis FARIÑAS DEL CERRO de m'avoir accueillie au sein du Laboratoire IRIT.

Je tiens à remercier tout particulièrement Monsieur Claude CHRISMENT pour m'avoir acceptée dans l'équipe SIG et m'avoir guidée dans le monde de la recherche à travers la direction de ma thèse et tous ses précieux conseils.

Je tiens à exprimer ma profonde gratitude à Madame Josiane MOTHE pour l'intérêt et la disponibilité qu'elle a manifestés à l'égard de mes recherches ainsi que pour son soutien et sa patience. Qu'elle soit ici assurée de mon très grand respect et du plaisir que j'ai à travailler avec elle.

Je souhaite exprimer toute ma reconnaissance à Madame Dominique RIEU et à Monsieur Gilles KASSEL pour l'honneur qu'ils me font en acceptant d'être les rapporteurs de ce mémoire ; leur lecture attentive et leurs remarques ont permis d'en améliorer la rédaction.

Par leurs conseils avisés, Monsieur Fionn MURTAGH et Monsieur Ollivier HAEMMERLE m'ont permis de préciser et d'améliorer mon travail. Je tiens à leur témoigner toute ma gratitude d'avoir bien voulu être examinateurs dans ce jury.

Je remercie Madame Françoise GENOVA avec qui j'ai eu le plaisir de travailler dans le cadre du projet Masses de Données en Astronomie et qui me fait l'honneur de participer à ce jury.

Ces trois années de recherche n'auraient pas été aussi riches sans les précieuses discussions que j'ai eues avec Madame Nathalie AUSSENAC-GILLES. Au-delà de ses relectures pertinentes qui ont largement contribué à l'amélioration de la qualité de ce mémoire, je tiens à la remercier de faire partie de mes modèles.

Je remercie Monsieur Didier BOURIGAULT de m'avoir permis d'utiliser le logiciel Syntex ainsi que de m'avoir consacré du temps et m'avoir fait bénéficier de son savoir-faire.

La partie évaluation de ce travail a été possible grâce à la patience et la disponibilité des astronomes du CDS, Pascal DUBOIS, Andrea PREITE MARTINEZ, Sébastien DERRIERE ainsi que de Soizick LESTEVEN qui ont rendu mes déplacements à Strasbourg enrichissants et agréables.

Mes remerciements vont également vers tous les membres de l'équipe SIG (Estella, Désiré, Ronan, Guillaume, Max, Olivier, Gilles) pour leur convivialité durant ces quatre dernières années. Plus particulièrement, un grand merci à Madame Florence SEDES d'avoir porté un intérêt particulier et pertinent à mes travaux.

Je ne voudrais pas oublier le personnel de l'irit (Agathe, Jean-Claude, Jean-Pierre, …) : merci pour votre bonne humeur et les services que vous m'avez rendus.

Je remercie également mes collègues du Département de Mathématique et Informatique de l'Université de Toulouse Le Mirail pour leur accueil chaleureux dans l'équipe pédagogique et la compréhension dont ils ont fait preuve en ce début d'année à l'emploi du temps chargé.

Je tiens également à remercier Madame Asuncion GOMEZ PEREZ ainsi que tous les membres de l'équipe Ontology Engineering Group de m'avoir chaleureusement accueillie à Madrid pendant trois mois. Plus particulièrement, je remercie Marie Carmen, Raul, Angel et Oscar qui m'ont montré ce qu'était la convivialité espagnole et qui ont rendu mon séjour mémorable.

Un grand merci à Laurent CARDONER qui a largement participé à la réalisation du prototype OntoExplo et qui, au fil de nos travaux, est devenu un ami.

Comment ne pas remercier « les filles », Asma BRINI et Karen SAUVAGNAT, qui m'ont supportée au quotidien pendant cette thèse, que ce soit par leur soutien dans les moments

de doute ou par leur présence dans les moments importants. Karen, je ferai tout aussi pour préserver notre amitié si précieuse. Asma, notre amitié est inestimable, tu m'as ouvert les yeux sur la différence et tu m'as montré que lorsqu'on se fixe un but, il faut s'y tenir quoi que la vie nous réserve.

Je remercie également mon « grand frère » Saïd KAROUACH, Benoît ENCELLE et Cédric BAUDRIT pour les bouffées d'air pur partagées autour des nombreux cafés qui ont rythmé ces dernières années.

Mes remerciements vont également vers mes amis de toujours, Mélanie, Bertrand, Jèf et Laura qui ont toujours été là pour partager les soucis, les joies et les moments de détente. Merci d'avoir accepté que je sois moins disponible cette dernière année, aussi importante qu'elle ait été pour certains…

Un grand merci à la famille Anton éparpillée aux quatre coins du monde qui, à travers son amitié, est certainement à l'origine de mon goût pour l'informatique et l'enseignement.

Je tiens à remercier profondément mes parents qui ont toujours été d'un soutien inconditionnel, que ce soit par leur disponibilité, leur générosité, la richesse de leur éducation (sans lacune !), les relectures du fond du désert argentin, les repas improvisés (la liste serait trop longue !)... Qu'ils soient ici en partie récompensés pour tout ce qu'ils m'ont donné.

Enfin, je voudrais remercier les quatre piliers qui m'ont aidée à tenir bon ces trois dernières années :

Ma sœur Claire à laquelle je me sens liée par une force magique qui, peu importe la distance qui nous sépare, m'aide toujours à garder la tête haute.

Fleur MOUGIN avec qui j'ai le bonheur de partager mon goût pour le Web Sémantique. Merci de m'avoir toujours tendu la main pour me sortir des rundopuntos que je fabrique si facilement dans mon pensamiento.

Loïc qui a tenu son contrat de confiance que ce soit sur les bancs du DEA ou du fond de sa Bretagne.

Et, last but not least, Anthony qui, par la force de nos différences, m'a aidée à gravir cette montagne que représente une thèse. Merci de m'avoir écoutée patiemment, d'avoir supporté les sautes d'humeur, les remises en cause et le silence studieux que j'ai imposé à la maison. A quand la prochaine montagne ?

Table des Matières

Introduction générale

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations. Afin de faciliter la mise en place de systèmes pouvant gérer de grandes quantités d'information et de s'adresser à un maximum d'utilisateurs, de nombreuses suppositions pragmatiques et simplifications ont été avancées dans la littérature [Jones 2004]. L'une d'elles a consisté à proposer des systèmes pouvant être utilisés par le plus grand nombre d'utilisateurs dans la plupart des cas [Spark Jones 1999]. Ainsi, le fonctionnement du noyau des Systèmes de Recherche d'Information (SRI) est indépendant du contexte. Les mécanismes et les méthodes qu'ils mettent en place sont les mêmes quel que soit le contexte de la recherche (documents composant le corpus, requêtes), l'utilisateur, son type de besoin en informations et l'usage qu'il souhaite faire de l'information retrouvée. Ces mécanismes se focalisent sur la représentation des documents et des requêtes soumises au système et leur mise en correspondance, mettant de côté la modélisation du contexte lié à l'utilisateur et sa recherche. Afin de combler ces lacunes et de proposer des systèmes répondant plus précisément au besoin utilisateur, le domaine de la RI contextuelle est apparu récemment comme une priorité [Allan 2003]. L'objectif de la RI contextuelle est de replacer l'utilisateur au cœur des modèles en rendant explicites certains éléments du contexte qui peuvent influencer sur les performances des systèmes. Le contexte fait référence aux connaissances relatives, aux intentions de l'utilisateur (tâche à accomplir, perception de la tâche, type d'information recherchée), à l'utilisateur lui-même (connaissance a priori, profil, culture), à son environnement (environnement matériel, historique des tâches), au domaine du besoin en information (nature du corpus, domaines abordés) et aux caractéristiques du système (représentation des documents, méthode d'appariement requête/document, interface de visualisation, stratégies d'accès à l'information). La prise en compte du contexte dans les SRI implique à la fois d'identifier puis de modéliser les différents aspects du contexte utiles pour la spécification du besoin de l'utilisateur et de les intégrer dans les méthodes et processus de RI. [Taylor 1968] dissocie deux paramètres distincts mais imbriqués par rapport au besoin en information. Le premier paramètre est le thème ou le sujet du besoin qui détermine sur quoi devra porter l'information recherchée. Le second paramètre relève de la tâche et de la situation dans laquelle se trouve l'utilisateur. Ce paramètre conditionne les raisons pour lesquelles l'information est recherchée et comment celle-ci sera utilisée. La majorité des SRI se focalise sur le premier paramètre, le thème [Freund 2005b]. De plus, cet aspect n'est pris en compte que partiellement puisque ces systèmes se contentent de rechercher dans les documents les termes donnés par l'utilisateur pour spécifier le thème du besoin.

Cette thèse vise à mieux prendre en compte deux types de connaissances liés au contexte : le thème du besoin en l'incluant dans son domaine et la tâche de l'utilisateur. Les ontologies sont un moyen de représenter la connaissance. Ces représentations de connaissances correspondent à « une spécification explicite et formelle d'une conceptualisation partagée » [Studer 1998]. Etant au cœur du Web sémantique pour ajouter une couche sémantique au Web actuel, elles font l'objet de nombreux travaux de recherche. Ces travaux s'attachent, d'une part, à définir des méthodologies et des techniques permettant leur élaboration à partir de textes et, d'autre part, à leur utilisation dans les systèmes d'information. Une ontologie fournit une référence pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets. Ceci est fait tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine [Aussenac 2004]. L'utilisation d'ontologies dans un modèle de RI a pour finalité de spécifier des

connaissances qui seront interprétables d'une part par l'utilisateur du système et d'autre part par le système lui-même.

Les ontologies dites « légères » contiennent des concepts et des relations entre concepts ainsi qu'un lexique permettant de référencer les concepts et les relations mais n'intègrent pas d'axiomes dans leur formalisation contrairement aux ontologies lourdes. Dans le cadre de la RI, l'utilisation des ontologies légères présente un niveau de formalisation suffisant pour mettre en place un nouveau type d'indexation, qualifié d'indexation sémantique, sur de grandes quantités de documents. L'indexation sémantique repose sur l'intuition selon laquelle le sens des informations textuelles (et des mots qui composent les granules) dépend des relations conceptuelles entre les objets du monde auxquels elles font référence, plutôt que des relations linguistiques et contextuelles trouvées dans leur contenu [Haav 2001]. L'indexation sémantique consiste à rechercher les concepts référencés dans les documents et à pondérer ces concepts en fonction de leur représentativité dans les documents. Les ontologies, pour être utilisées dans le cadre de la RI, doivent avoir une large composante lexicale afin que les termes référençant les concepts puissent être retrouvés dans les documents.

Dans ce contexte, nous proposons un modèle à base d'ontologies dont l'objectif est de représenter les deux aspects du contexte que nous avons indiqués précédemment : l'aspect lié à la tâche de recherche et celui lié au thème du domaine. Notre cadre d'étude est donc celui de bases documentaires d'un domaine spécifique. Notre modèle s'appuie sur des ontologies de domaine. Par opposition aux ontologies génériques, les ontologies de domaine se limitent à représenter la connaissance d'un domaine particulier. Notre choix est motivé par le fait que les ontologies de domaine restreignent l'interprétation des concepts qu'elles définissent au contexte spécifié par le domaine. Ceci a l'avantage de limiter l'ambiguïté des termes définis dans l'ontologie pour référencer les concepts facilitant ainsi leur détection dans les documents. Pour modéliser les deux aspects du contexte, notre modèle repose sur deux ontologies de domaine. Une première ontologie spécifie et structure les objets du thème traités dans les documents ainsi que leurs relations (ontologie du domaine du thème). La seconde spécifie les données qui intéressent l'utilisateur par rapport au type de tâche qu'il accomplit (ontologie du domaine de la tâche). L'intégration du modèle dans le SRI est au cœur de notre thèse et intervient dans deux phases du processus de recherche. Par la proposition d'un mécanisme d'indexation sémantique reposant sur les deux ontologies, il est intégré à la phase de représentation des documents. L'originalité de notre approche repose sur le fait que les deux aspects du contexte sont liés par l'utilisation d'éléments communs aux deux ontologies. De plus, le modèle est intégré à la phase d'accès aux documents via la navigation dans les ontologies. Une autre originalité de notre approche est que cette navigation repose sur deux niveaux d'accès à l'information. Le niveau concept donne à l'utilisateur une vue globale sur la collection de documents et sur la connaissance associée, alors que le niveau instance donne un accès aux informations spécifiques contenues dans les documents.

L'utilisation d'ontologies en RI pose une autre problématique qui est la réutilisation de la connaissance déjà modélisée. En effet, de nombreuses ressources terminologiques (comme les thésaurus) ou conceptuelles existent dans différents domaines. Nous avons étudié la réutilisabilité de telles ressources selon deux perspectives : le choix d'une ontologie légère en fonction de son adéquation au corpus à indexer et l'élaboration d'une ontologie légère à partir d'un thésaurus normalisé et d'un corpus de référence. Une originalité de nos travaux concernant l'évaluation de l'adéquation réside dans la prise en compte de l'ensemble des relations définies dans les ontologies et non pas seulement des relations taxonomiques. L'adéquation intègre l'ensemble de

la connaissance contenue dans l'ontologie légère. Concernant l'élaboration d'une ontologie légère à partir d'un thésaurus, une de nos contributions est de proposer un mécanisme semi-automatique pour capturer la connaissance représentée dans le thésaurus et la mettre à jour à partir de documents de référence.

L'ensemble des propositions a été validé dans le cadre d'un projet en coopération avec des astronomes (Masses de Données en Astronomie). Nous avons pu ainsi évaluer un certain nombre des techniques que nous proposons. Un prototype illustre également l'apport de nos contributions.

Le mémoire est organisé comme suit.

Les chapitres 1, 2 et 3 présentent l'état de l'art relatif aux domaines en lien avec nos travaux.

Le **chapitre 1** introduit le domaine de la RI contextuelle et l'utilisation de connaissances sous-jacentes dans ce domaine. La notion de connaissance est située par rapport à ses définitions dans le domaine de l'Ingénierie des Connaissances et par rapport aux différentes représentations qui sont associées aux ontologies.

Le **chapitre 2** présente différents travaux de la littérature en lien avec la construction d'ontologies à partir de textes. Il précise également les travaux relatifs à la transformation d'un thésaurus en une ontologie.

Le **chapitre 3** décrit l'utilisation des ontologies en RI à partir des différents processus dans lesquels elles sont intégrées. Après avoir présenté les travaux visant à analyser et à mesurer la réutilisabilité des ontologies en général, et leur réutilisabilité en RI en particulier, nous présentons leur utilisation aux niveaux de l'indexation des documents et de l'accès aux documents.

Les chapitres 4, 5 et 6 décrivent nos contributions théoriques.

Le **chapitre 4** décrit le modèle de représentation du contexte que nous proposons. Il repose sur les notions d'ontologie du domaine de la tâche et du domaine du thème qui sont décrites en détail. Ce chapitre comprend également notre proposition d'intégration du modèle dans un processus de RI par l'indexation sémantique et l'accès aux granules documentaires à partir des deux ontologies.

Le **chapitre 5** présente la méthode de transformation d'un thésaurus en une ontologie légère que nous proposons. Cette méthode repose sur différentes étapes qui permettent d'extraire la connaissance contenue explicitement dans le thésaurus (lexique de l'ontologie) et implicitement (concepts et relations) ainsi que de la mettre à jour en s'appuyant sur un corpus.

Le **chapitre 6** expose la méthodologie que nous avons définie pour évaluer l'adéquation entre un corpus et une ontologie légère. Il précise également des méthodes permettant de la mettre en œuvre.

Le **chapitre 7** présente le cadre applicatif de nos travaux ainsi que les expérimentations mises en place pour évaluer nos propositions.

Le **chapitre 8** correspond au prototype que nous avons développé pour valider nos propositions.

Partie 1

Etat de l'art

Chapitre 1

Modélisation du contexte d'une recherche à partir de représentations de la connaissance

1 Introduction

De manière générale, l'utilisation de connaissances en informatique a pour but de ne plus faire manipuler en aveugle des informations à la machine mais de permettre un dialogue, une coopération entre le système et les utilisateurs (système d'aide à la décision, système d'enseignement assisté par ordinateur, recherche d'information). Pour cela, le système doit avoir accès non seulement aux termes utilisés par l'être humain mais également à la sémantique qui leur est associée, afin qu'une communication efficace soit possible. Les ontologies visent à représenter cette connaissance en étant à la fois interprétables par l'homme et par la machine.

L'Ingénierie des Connaissances (IC) est une branche de l'intelligence artificielle axée sur la connaissance. Les principales préoccupations de ce domaine sont l'acquisition, la modélisation, le stockage et la consultation de connaissances, le raisonnement automatique sur les connaissances stockées et la modification des connaissances stockées. Lorsque les connaissances à construire sont issues de documents, l'IC s'appuie sur des méthodologies développées dans le domaine de la linguistique et du traitement automatique des langues pour assurer une compréhension des contenus des documents considérés.

Parallèlement, la Recherche d'Information (RI) est une activité dont la finalité est de mettre en regard des informations et un utilisateur. C'est une activité par laquelle un utilisateur accède à un granule d'information (un ensemble de documents, un document, une partie de document, un composant XML, une donnée) à partir d'un besoin qu'il spécifie. Les systèmes de RI (SRI) développés depuis le début des années 50 reposent essentiellement sur des approches statistiques et des approches linguistiques de bas niveau. Ces approches prennent uniquement en compte le niveau lexical, parfois le niveau syntaxique, du contenu textuel des granules afin d'identifier les mots permettant de retrouver les granules répondant aux besoins de l'utilisateur. Un enjeu actuel de la RI, comme du Web avec le Web Sémantique, est de s'appuyer sur des connaissances pour enrichir les systèmes en apportant une couche sémantique.

Dans le cadre de la RI, la problématique posée est de doter le SRI de connaissances lui permettant d'être un intermédiaire entre l'utilisateur et les granules d'information. Ce rôle d'intermédiaire se joue entre un utilisateur dont le système doit être capable d'interpréter les besoins et des granules d'information dont le système doit pouvoir interpréter le contenu. En d'autres termes, l'utilisation de connaissances par un SRI doit lui permettre de connaître le contexte de la recherche. D'un côté, le système doit être capable d'inférer les intentions de l'utilisateur en fonction de la tâche visée et, de l'autre, d'expliciter le contenu d'un granule d'information.

L'objectif de ce chapitre est de positionner l'utilisation de connaissances par rapport au processus de RI en analysant la notion de connaissance du point de vue de l'IC.

Dans un premier temps (section 2), nous présentons la notion de contexte en RI et la connaissance sous-jacente. Nous définissons dans la section 3 la notion de connaissance par l'explicitation de son lien avec les notions d'informations et de données, par la définition de ses différentes caractéristiques ainsi que l'utilité d'en réaliser des représentations. Dans la section 4, nous détaillons différents types de représentations de connaissances associés à la notion d'ontologie et nous analysons dans quelle mesure les ressources de connaissances peuvent être intégrées aux processus de RI.

2 Connaissances sur le contexte d'une recherche d'information

L'exploitation du contexte en RI repose sur deux fondements [Johnson 2003]. Dans le domaine du langage et de la communication, le contexte est utilisé pour appréhender le sens des mots. Il permet de désambiguïser le sens des unités sémantiques en fonction du domaine, de la

discipline et du contexte linguistique (document, phrase …) auxquels elles appartiennent. De plus, le contexte permet de déterminer l'action sociale dans laquelle se place un individu. Il est en effet plus facile de modéliser le comportement des utilisateurs appartenant à un groupe restreint et identifié que de déterminer les caractéristiques propres à l'ensemble de la population. Comme le souligne [Freund 2005b], ces deux aspects sont fondamentaux en RI.

Plus généralement, la notion de contexte recouvre en RI différents aspects. Dans [Jarveling 2004], un modèle analytique de la RI est présenté (cf Figure 1 .1).

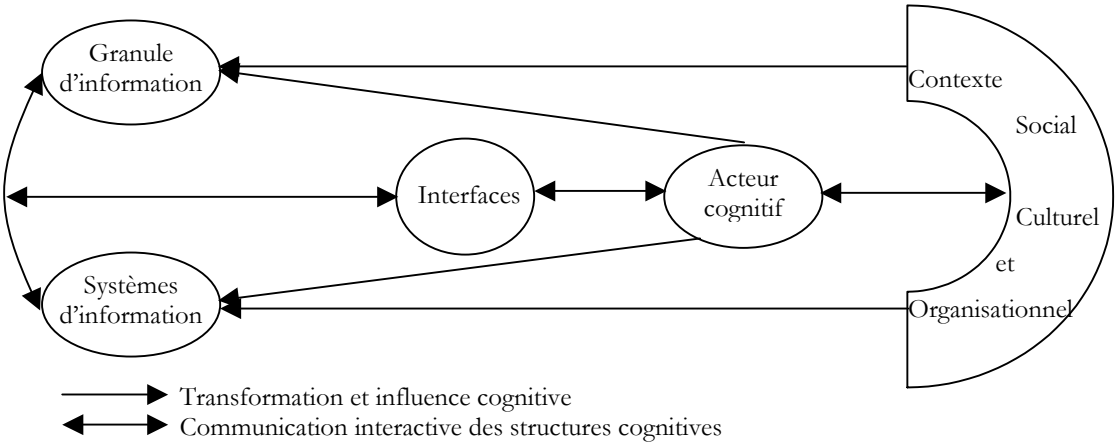


Figure 1.1 Modèle cognitif de la recherche d'information adapté de [Jarveling 2004]

Chacun des acteurs est représenté (nœud du graphe) ainsi que leurs liens. Les acteurs cognitifs (comme les utilisateurs recherchant une information) sont entourés de différents acteurs de contexte (granules d'information, systèmes d'information, interface, contexte social, culturel et organisationnel). Tous ces acteurs sont en interaction. Les granules d'information correspondant aux éléments gérés par le système d'information peuvent être très variés (collections de documents, documents, parties de documents, phrases, données). Le système d'information lui-même repose sur différents modèles et méthodes (modèle de représentation de l'information, méthode de mise en correspondance), en lien avec les objets d'information gérés et l'interface de manipulation. L'acteur cognitif de son côté est influencé dans sa recherche par le contexte culturel et social dont il dépend ; les objets auxquels il s'intéresse en dépendent. Le contexte du système dépend à la fois de l'information gérée et des spécificités de l'utilisateur. L'interface, qui permet de faire la correspondance entre l'utilisateur et la collection d'objets traités et restitués par le système, doit s'adapter en fonction des contextes (types d'information manipulée, utilisateur, technologie de RI utilisée).

Les systèmes de RI actuels visent à satisfaire la majorité des utilisateurs dans la plupart des cas [Spark Jones 1999]. La variabilité des contextes qui viennent d'être présentés les rend difficiles à modéliser. Pour cette raison, bien que le contexte soit omniprésent, les systèmes de RI n'en intègrent que certains aspects. Les travaux menés dans le cadre du workshop IRiX (Information Retrieval in conteXt) en 2005[1] ont conclu que quatre dimensions principales du contexte pouvaient être retenues : les granules d'information, la tâche, l'utilisateur, le système. Les sections suivantes présentent comment les différents aspects du contexte sont ou ne sont pas intégrés dans les systèmes.

[1] http://www.dcs.gla.ac.uk/IRiX/

2.1 Contexte et granules d'information

2.1.1 Représentation du contenu des granules d'information

Les index jouent un rôle primordial en RI en définissant les descripteurs (mots ou groupements de mots) à partir desquels l'information contenue dans les granules est représentée. Le langage d'indexation est donc un langage artificiel, c'est-à-dire construit à l'aide d'un ensemble de règles données, servant à la représentation abrégée du contenu d'un document [Rivier 1990]. Dès lors, l'indexation consiste à détecter les termes les plus représentatifs du contenu du document.

Deux idées très simples ont rendu possible l'indexation automatique du contenu textuel de documents : considérer les mots qui apparaissent dans les granules et compter leurs occurrences [Zipf 1949], [Roberston 1976]. Les recherches dans le domaine ont montré qu'il est essentiel pour une indexation efficace et donc pour une recherche efficace de prendre en compte la distribution de ces mots dans les documents car les fréquences relatives auxquelles les mots apparaissent ou co-occurrent permettent de déterminer les thématiques et leur signification dans les textes [Spärk Jones 2003].

D'ailleurs, les modèles de RI se sont attachés à rendre la meilleure représentation possible du contenu de la collection de documents (indexation) pour leur mise en correspondance avec la requête. A l'inverse des premiers systèmes automatiques, il s'agit de ne plus considérer que les termes issus des documents sont indépendants [Rijsbergen 1979]. Il s'agit là d'une forme de connaissance sur les termes et leur utilisation. Plusieurs méthodes ont été introduites.

La représentation des documents par des radicaux, plutôt que par les termes tels qu'ils apparaissent dans les documents, est une première démarche permettant une meilleure représentation des contenus et prenant en compte la dépendance entre termes. Cette radicalisation peut s'appuyer sur différentes stratégies : troncatures simples [Denjean 1989], suppression des suffixes [Porter 1980], utilisation de connaissances linguistiques [Savoy 1993]. Il s'agit ici de considérer les différentes formes lexicales d'un terme comme équivalentes. La majorité des SRI actuels reposent sur une représentation par termes radicalisés.

Cependant les mots retenus par l'indexation peuvent être ambigus. Les descripteurs peuvent en fait se rapporter à des termes ayant plusieurs sens et donc ne pas indiquer clairement la thématique abordée dans le document. Différents descripteurs peuvent également se rapporter à une même notion dans le cas où les mots choisis sont synonymes. L'index est alors surchargé par des éléments représentant la même information. D'autre part, la recherche peut échouer si les termes de la requête n'y apparaissent pas. L'utilisation de connaissances dans le but d'aider le SRI à interpréter le contenu des documents permet d'accéder à la sémantique associée aux mots issus du contenu. Cette sémantique repose sur l'identification des concepts et des relations entre ces concepts de manière à établir clairement les notions, les termes et les objets associés aux mots issus des textes des granules. Une représentation de la connaissance abordée dans le granule permet au système de prendre en compte la sémantique sous-jacente aux mots qui composent le contenu des granules. De façon générale, une analyse statistique permet l'extraction de descripteurs des granules d'information mais pas leur compréhension. [Nazarenko 2004] définit la « compréhension de texte » comme « être capable de modifier sa représentation du monde en fonction des informations véhiculées par le texte ». Ceci signifie que l'utilisateur ou le système possède un ensemble de connaissances préalables et que la « compréhension du texte » lui permet de modifier cette connaissance en ajoutant, supprimant ou modifiant la connaissance qu'il avait déjà. De la même façon, l'enjeu des activités documentaires est décrit dans le projet ASSTICCOT comme visant à permettre que des connaissances produites par un auteur engendrent des connaissances "nouvelles" c'est-à-dire différentes, pour les utilisateurs [ASSTICCOT 2003]. Typiquement, les connaissances diffusées dans un brevet d'invention vont permettre de produire

d'autres connaissances pour les utilisateurs (connaissances se traduisant par un positionnement stratégique par exemple dans le domaine de la veille). Représenter des documents d'un domaine à partir de la connaissance issue de ce domaine, peut permettre de mettre à jour cette connaissance par le traitement des documents. L'opération effectuée par le système se rapproche alors de la compréhension des documents. Fournir à l'utilisateur la ressource de connaissances permettant le traitement par le système peut, de plus, l'aider dans sa compréhension des documents.

Certains auteurs affirment que l'indexation peut être considérée comme une forme d'acquisition de connaissances sur le contenu documentaire [Dachlet 1990]. Or, comme nous le décrirons dans la section 2, la connaissance est une information active et interprétée par l'homme. Dans le cas des index, l'information extraite a pour but de représenter l'information du granule et de permettre d'établir une correspondance entre son contenu et le besoin de l'utilisateur. Cette information n'est pas active dans le sens où le seul processus dans lequel elle s'inscrit est la mise en correspondance. Elle ne permet ni d'inférer de nouvelles connaissances, ni d'être interprétée par l'homme pour qu'il mette à jour ou ne complète ses connaissances.

Le traitement des documents doit prendre en compte non seulement leur contenu textuel mais aussi les méta-données qui peuvent être associées aux documents.

2.1.2 Méta-données associées aux granules d'information

Les méta-données sont des données factuelles qui contiennent de l'information sur l'information des granules. Plus précisément, c'est un ensemble structuré d'informations décrivant une ressource. Elles sont associées aux ressources sans ambiguïté comme, par exemple, le nom des auteurs, la date de publication, les mots clés choisis pour indexer le document… Les ressources étant généralement partagées, plusieurs standards ont été définis pour permettre leur description à l'aide des méta-données.

Le Dublin Core (http://www.dublincore.org/) définit un ensemble de 15 méta-données relatives

- au *Contenu*: Titre, Description, Sujet, Source, Couverture, Type, Relation,
- à la *Propriété intellectuelle*: Créateur, Contributeur, Editeur, Droits,
- à la *Version*: Date, Format, Identifiant, Langage.

Les méta-données associées par le Dublin Core sont considérées comme descriptives car elles sont externes aux contenus même des documents et elles indiquent comment le granule a été créé [Baeza-Yates 1999]. D'autres types de méta-données comme les méta-données associées par le système Medline (http://medline.cos.com/) sont relatives aux contenus même des granules. Des méta-données à propos de maladies, symptômes ou anatomies sont associées par ce système à des articles de médecine.

RDF (Resource Description Framework) [Lassila 1999] est un moyen d'encoder, d'échanger et de réutiliser des méta-données structurées. Les méta-données peuvent aussi bien être descriptives que relatives aux contenus des granules. RDF est décrit en détail dans la section 4.3 de ce chapitre.

Pour être interprété par un système, il est nécessaire qu'une sémantique soit associée à ces méta-données. Le système doit être capable d'interpréter le rôle de la méta-donnée dans la représentation du document. De plus, il doit être capable d'interpréter les liens entre différentes méta-données associées aux documents. Une ontologie permet de spécifier la connaissance nécessaire au système pour interpréter le rôle sémantique des méta-données.

Peu de systèmes intègrent à la fois des descripteurs liés aux contenus des granules et aux méta-données.

2.1.3 Représentation de la requête et reformulation

A l'autre bout de la chaîne de RI, les requêtes sont considérées. La représentation de la requête se limite à l'ensemble des termes issus de la formulation de la requête par l'utilisateur. Cependant, les mécanismes de reformulation de requêtes permettent d'améliorer cette représentation à partir de connaissances extraites des contenus des granules ou de ressources externes.

Un premier type d'approches repose sur l'analyse globale de la collection de documents considérée. L'extraction des liens de co-occurrence entre termes des documents, calculés de façon statistique [Harper 1978] ou faisant intervenir des connaissances linguistiques [Grefenstette 1992], l'extraction de groupes de mots [Pohlmann 1997] et celle des liens contextuels entre termes [Bruandet 1983], [Véronis 1989], font partie de cette catégorie. Les informations ainsi extraites sont généralement utilisées pour reformuler automatiquement une requête par ajout des termes liés aux termes initialement présents dans la requête. Les termes ainsi ajoutés sont issus des documents et permettent donc une meilleure adéquation entre le besoin d'information et la collection.

L'ajout de termes issus de ressources terminologiques est une autre méthode de reformulation de requête. Dans ce cas, les termes liés aux termes initiaux de la requête sont extraits de thésaurus [Jarvelin 1996] ou de ressources telles que WordNet [Voorhees 1993], [Mandala 1999].

Enfin, le principe de réinjection de pertinence [Rocchio 1971], [Harman 1992] vise également à reformuler une requête initiale pour qu'elle corresponde mieux au contenu de la collection. Le principe est le suivant : l'utilisateur soumet sa requête initiale, le système restitue un premier ensemble de documents que l'utilisateur doit juger (pertinent, non pertinent). La connaissance de la pertinence des documents initialement restitués est utilisée pour sélectionner des termes à ajouter à la requête initiale (les termes qui sont caractéristiques de la pertinence en quelque sorte), voire décider des termes à éliminer de la requête [Rocchio 1971], [Salton 1990]. Cette méthode repose donc sur l'hypothèse que les documents pertinents se ressemblent [Rijsbergen 1979]. Des améliorations de 20 à 30% de la précision moyenne ont été mesurées [Harman 1992]. L'ambition du « tout automatique » a fait dire que cette méthode impliquait un dialogue système / utilisateur trop lourd. Ainsi, cette méthode a été automatisée. Pour éviter la lourdeur du mécanisme de jugement de pertinence des documents initialement restitués, la réinjection de pertinence aveugle prend en compte non pas la pertinence utilisateur, mais la pertinence système. Dans cette méthode, les premiers documents restitués par rapport à la requête initiale sont considérés comme pertinents. Seule une pertinence positive est alors considérée. Des études ont montré que cette approche permettait d'améliorer les résultats par rapport à une recherche simple [Harman 2000]. Parallèlement, l'utilisation du contexte local des termes de la requête dans les documents a également été étudiée [Xu 2000].

2.1.4 Domaine

Les portails thématiques ou les librairies électroniques, en se focalisant sur un thème ou un usage, considèrent l'aspect culturel et social du contexte (cf Figure 1.1). Par exemple, les portails de l'ACM[2], de CiteSeer[3], etc. ont pour cible les scientifiques. Le contexte est donc partie intégrante de la collection, soit par son aspect thématique (collection en Médecine comme

[2] http://portal.acm.org/

[3] http://citeseer.ist.psu.edu/

MedLine[4] par exemple), soit par le type de documents qu'elle contient (publications scientifiques). L'utilisateur peut être plus confiant sur l'adéquation des documents qu'il va retrouver.

La notion de domaine peut être également abordée à travers la représentation des documents par un langage contrôlé. Par exemple, l'utilisation de la hiérarchie thématique MeSH (Medical Headings) pour représenter les documents de MedLine impose de fait le vocabulaire utile pour la recherche (celui qui permettra de retrouver des documents s'il est utilisé dans une requête) [Hearst 1997]. Le même type d'approche est utilisé dans le système IRAIA [Englmeier 2003] gérant des documents économiques. L'approche multi-facette retenue ici permet de représenter les documents selon différentes hiérarchies de concepts, chacune correspondant à un aspect du domaine (type d'entreprise, pays, indicateurs économiques).

L'indexation de collections à partir d'un vocabulaire de domaine présente de plus les avantages suivants :

- Aider l'utilisateur à formuler sa requête. En présentant le vocabulaire du domaine à l'utilisateur, il est possible de le guider dans le choix des termes de sa requête. Il a été montré qu'un thésaurus permet à l'utilisateur de construire une conceptualisation de ce qu'il est en train de chercher et peut aider à la formulation des requêtes dans le cas de la RI ad-hoc [Baeza-Yates 1999]. Cependant, de nombreuses limites ont été trouvées à l'utilisation de thésaurus [Baeza-Yates 1999] car, d'une part, leur construction est orientée terminologie et ne capture que les termes d'un domaine et, d'autre part, les relations entre termes restent limitées sémantiquement. L'utilisation de ressources conceptuelles permet de combler ces lacunes,
- Faciliter la RI au sein de collections hétérogènes en indexant tous types de documents à partir des mêmes termes.

La connaissance associée à un domaine peut être représentée de façon plus formelle au travers d'une ontologie. Pour un utilisateur, accéder par une ontologie à la connaissance à partir de laquelle l'information d'un corpus a été indexée peut lui permettre de spécifier son besoin et les lacunes de sa connaissance par rapport à l'information qui lui est disponible. D'autre part, la représentation des granules d'information à partir d'une ontologie peut définir un vocabulaire contrôlé (termes et concepts) à partir duquel l'utilisateur spécifiera son besoin. La description du besoin correspond, dans ce cas-là, aux caractéristiques des granules car elles ont été indexées à partir des mêmes ressources. Cet aspect, au cœur de notre thèse, sera développé en détail dans le chapitre 3.

2.2 Contexte et utilisateur

Différents facteurs ont été proposés pour représenter le contexte de l'utilisateur [Belkin 2004] :

- la familiarité de l'utilisateur avec le domaine relatif à sa recherche [Kelly 2002],
- l'expérience de l'utilisateur dans l'utilisation du ou des systèmes de RI,
- les documents déjà connus de l'utilisateur,
- le type des documents recherchés [Rauber 2001], [Freund 2005a],
- l'objet de la recherche,
- la tâche dans laquelle s'inscrit la recherche,

[4] Base de données en médecine, http://www.chu-rouen.fr/documed/medline.html

- les autres activités de l'utilisateur pendant sa recherche.

L'utilisateur s'engage dans une recherche d'information parce qu'il a un manque d'information. Cependant, l'utilisateur a une idée plus ou moins définie des lacunes de ses connaissances et donc de son besoin en information. La première difficulté à laquelle doit faire face un SRI est que le besoin en information est interne à l'utilisateur. L'utilisateur juge en effet les éléments qui lui sont retournés par rapport à l'interprétation de son besoin et non pas par rapport à l'ensemble des granules à sa disposition et susceptibles de l'intéresser dans la collection [Turtle 1991].

Ainsi, le projet Profildoc [Lainé-Cruzel 1999] permet de filtrer les documents retrouvés par rapport au profil de l'utilisateur. Ce profil utilisateur contient des informations telles que le niveau éducationnel, le champ disciplinaire (sciences de l'information, agronomie…), le type de recherche (recherche généraliste ou pointue)… Ce profil est utilisé afin d'identifier au sein des documents retrouvés ceux qui correspondent au profil de l'utilisateur et ainsi réduire le nombre de documents en éliminant ceux qui ne seraient pas pertinents pour lui.

De la même façon, la tâche « Hard » de TREC (trec.nist.gov) s'intéresse à l'étude de l'impact de certains de des facteurs liés à l'utilisateur sur les performances d'un système de RI [Allan 2003a]. Plus particulièrement, en 2003, les besoins d'informations comprenaient, en plus des champs traditionnels des « topics » TREC (titre, description, narration), les méta-données :

- familiarité avec le thème,
- type de documents souhaités,
- objet de la recherche,
- spécification géographique sur les documents recherchés.

En 2004, les méta-données ont été simplifiées pour ne retenir que :

- la connaissance sur le thème,
- le type de documents recherchés,
- la spécification géographique (le document est relatif aux Etats-Unis ou non).

Dans cette tâche, les participants doivent d'abord fournir les documents retrouvés sans prendre en compte les méta-données associées à la connaissance de l'utilisateur ; dans un deuxième temps, les participants fournissent les résultats qu'ils obtiennent lorsque leurs systèmes intègrent ces informations sur le contexte. L'influence du contexte est donc mesurée en comparant les premiers résultats avec les seconds.

Enfin, la recherche d'information collaborative correspond à un autre type de processus mis en œuvre pour considérer l'utilisateur. Cette technique vise à permettre à des utilisateurs de bénéficier de jugements de pertinence émis par d'autres utilisateurs supposés partager le même profil. Ces approches reposent essentiellement sur le contenu des informations recherchées. De façon similaire, les systèmes de recommandation visent à optimiser la recherche d'information en proposant automatiquement à l'utilisateur de nouveaux documents au regard de ses besoins exprimés ou de ses actions. Une étude de ces systèmes de recommandation appliqués au contexte d'Internet peut être trouvée dans [Montaner 2003]. A la frontière entre les outils de recommandation et de recherche d'information collaborative, Easy-DoR [Chevalier 2002] propose d'utiliser les utilisateurs comme source d'information pour la recommandation, et ce, au travers des documents qu'ils visitent sur le Web. Par ailleurs, ce système propose un type de filtrage collaboratif au travers de jugements de pertinence déduits de l'organisation des documents (signets) que chaque individu possède.

2.3 Contexte et tâche

Une tâche est définie comme « une activité réalisée pour atteindre un but » [Vakkari 2003]. Une tâche de recherche intervient quand l'utilisateur est en manque d'information. Pour étudier les besoins d'information, Dervin [Dervin 1992] propose, une méthode qui emploie la métaphore « *situation-gap-use* » selon laquelle tous les besoins d'information viennent d'une lacune dans les connaissances d'un individu ; cette lacune entraîne une situation spécifique à laquelle l'individu peut remédier par différentes tactiques. Le but de la recherche détermine le type d'information que l'utilisateur sollicite et l'utilisation qu'il souhaite faire de cette information.

Les différents programmes internationaux d'évaluation proposent une panoplie des tâches de recherche d'information (TREC, INEX, CLEF). Nous décrivons celles qui sont les plus communément développées dans les systèmes actuels.

La RI ad-hoc vise à restituer (tous) les documents pertinents (et seulement ceux là) par rapport à un besoin d'information formulé sous forme de requête par un utilisateur. La plupart des SRI fonctionnent avec une interface qui permet à l'utilisateur de formuler son besoin en information à partir d'une requête. Le système présente ensuite à l'utilisateur le résultat de la recherche sous forme d'une liste de références vers les documents retrouvés. S'il s'agit de la tâche la plus connue, d'autres tâches de RI existent. L'utilisateur peut souhaiter ne consulter que les documents ou granules nouveaux en rapport avec son besoin d'information [Soboroff 2003] ou filtrer les documents par rapport à un profil de recherche [Roberston 2002]. Il peut vouloir une réponse à une question précise [Voorhees 2004] (systèmes questions-réponses) ou en rapport avec un cadre spécifique comme la génomique [Hersh 2004]. La recherche sur le Web [Hawking 1999] correspond à une tâche spécifique dans la mesure où la présence de liens hypertextes peut modifier l'idée de l'utilisateur sur son besoin d'information. La recherche multilingue ou crosslingue [Jones 2000] correspond à un autre type de tâche. Alternativement, l'utilisateur peut vouloir explorer une collection de documents pour les classifier [Sebastiani 2006] ou pour découvrir des informations non implicitement présentes dans les documents comme dans une activité de veille [Chrisment 2006]. La RI est également une activité qui est intégrée dans de nombreuses tâches comme l'apprentissage pédagogique (e-learning), la gestion de la mémoire d'entreprise, etc.

La nature des différentes tâches de RI est diverse et implique des traitements de l'information adaptés à chacun des objectifs qu'elles doivent atteindre.

2.4 Bilan : représentation des connaissances en RI

Le contexte de la RI relève de plusieurs aspects, chacun d'eux pouvant prendre des formes variées. L'intérêt de mieux représenter la connaissance implicite ou explicite liée à ce contexte est d'aboutir à des systèmes qui permettront de mieux répondre aux besoins des utilisateurs [Allan 2003b]. Ainsi, un des challenges à long terme du domaine de la RI est de proposer des modèles qui intégreraient l'ensemble de ces aspects dans un modèle unique. Notre thèse s'inscrit dans ce contexte. Dans ce travail, nous proposons un modèle qui prend en compte deux aspects du contexte en RI : le domaine de la tâche et le domaine du corpus.

Pour mener à bien ce travail, il nous a paru que les ontologies étaient un des meilleurs moyens d'ajouter une couche sémantique aux SRI par rapport à ces deux aspects. En effet, les ontologies permettraient aux SRI de s'appuyer sur de la connaissance pour interpréter le sens des mots contenus dans les granules d'information. Accéder aux sens des mots, aux concepts sous-jacents, aux relations sémantiques entre concepts permettrait aux systèmes d'établir une meilleure interprétation du contenu des granules qu'ils ont à gérer. Une pré-condition est cependant nécessaire pour prendre en compte les ontologies : celles-ci doivent intégrer une forte composante lexicale afin de permette l'interprétation des mots des documents. Ces ontologies pourraient de plus apporter de la connaissance utile à d'autres niveaux de compréhension. Elles

laissent entrevoir une meilleure gestion et interprétation des méta-données associées aux granules en permettant aux systèmes de capturer la sémantique qui leur est associée. Elles peuvent également servir aux systèmes à mieux interagir et comprendre la tâche de l'utilisateur à laquelle ils doivent répondre. Ceci devient possible en donnant au système accès à une représentation de la connaissance contenue dans le corpus documentaire et dans la tâche et en fournissant des mécanismes d'accès aux granules adaptés. Finalement, les ontologies peuvent permettre une meilleure spécification du rôle des ressources et connaissances dans la réalisation des différentes tâches de RI. Telles sont les hypothèses qui ont été à la base de ce travail de thèse.

3 Qu'est-ce que la connaissance ?

Définir la connaissance en soi relève de la philosophie. Le propos de cette section n'est pas de répondre à une telle question mais de caractériser la connaissance, notre cadre de réflexion étant l'informatique. Nous entendons ici par informatique, non pas « la science des ordinateurs », mais « la science du traitement de l'information ».

Nous situerons tout d'abord la notion de connaissance par rapport aux différentes notions auxquelles elle est associée dans le domaine de l'informatique. Nous définirons ensuite ces différentes caractéristiques et nous expliciterons l'intérêt d'en réaliser une représentation pour permettre sa manipulation.

3.1 De l'information à la connaissance

Il convient tout d'abord de caractériser la connaissance par rapport à plusieurs termes auxquels elle est abusivement assimilée. Même s'il n'existe pas de frontières clairement établies entre les notions de donnée, information, processus et connaissance, chacune de ces notions joue un rôle propre en fonction de son niveau d'entrée dans un processus d'action d'un système informatique [Charlet 2002].

La donnée est le moins porteur de sens de tous ces termes. Tout instrument informatique et technologique crée de l'accumulation de données. Les données ne sont ni vraies, ni fausses, ni significatives à moins d'être récupérées, représentées et réinterprétées. Elles sont transmises à un système ou un programme qui les traite, les modifie et les fait évoluer.

Toute information est issue de données qui sont structurées pour constituer une information. L'information fait référence aux « messages » qui peuvent être restitués par le système et à l'usage des données. Les données deviennent informations quand elles prennent un sens soit pour le système soit pour l'utilisateur.

L'information, constituée des données, devient connaissance à partir du moment où elle sert de fondement à une inférence, au déclenchement d'un processus [Lame 2002]. Une inférence est définie par Kasyer [Kasyer 1997] comme « une façon générique de désigner l'ensemble des mécanismes par lesquels des entrées (perceptives ou non) sont combinées à des connaissances préalables afin d'obtenir des comportements élaborés ».

3.2 Caractéristiques de la connaissance

Tâchons maintenant de caractériser la connaissance à partir de définitions et de travaux issus de la littérature dans le domaine de l'IC.

> *« Une connaissance est la capacité d'exercer une action pour atteindre un but. »* [Bachimont 2004].

> *« Il n'y a présomption de connaissance que si la faculté d'utiliser des informations à bon escient est attestée. »… « Tandis que les informations sont exploitées par des processus sans pouvoir modifier leur déroulement, les connaissances sont des données qui influencent le déroulement de processus. »* [Kayser 1997].

> *« La connaissance est l'information organisée qui est applicable à la résolution de problèmes. »* [Wolf 1990].
>
> *« La connaissance inclut des restrictions implicites et explicites entre objets ainsi que des opérations et des relations, qui permettent de définir des heuristiques générales et spécifiques comme les procédés d'inférences liés à la situation à modéliser. »* [Sowa 1984].

Information active

Les connaissances sont des informations actives, dans la mesure où elles peuvent influencer le déroulement d'un processus, produire de nouvelles informations ou permettre de prendre des décisions [Furst 2004]. La connaissance est définie dans un cadre bien précis et prend sa signification dans le contexte de son utilisation. On ne peut pas parler de connaissance a priori [Charlet 2002].

Interprétée par l'homme

On ne peut parler de connaissance qu'à partir du moment où l'information manipulée par le système prend un sens pour l'utilisateur, c'est-à-dire qu'il peut établir un lien avec cette information et celle qu'il possède déjà [Charlet 2002].

Outil informatique : support de sa genèse et de sa mémorisation

L'informatique permet la mémorisation et la genèse de la connaissance [Charlet 2002]. En effet, les outils et les supports de stockage informatiques permettent à l'homme de constituer des connaissances, de les accumuler et de les faire évoluer.

Théorique ou pratique

Bachimont distingue la connaissance par rapport à son caractère théorique et son caractère pratique [Bachimont 2004]. La connaissance pratique se réfère à des actions associées à une activité dans le monde matériel. Elle permet la modification physique et matérielle du monde. Elle renvoie au savoir faire. La connaissance théorique quant à elle correspond à une activité non pas dans le monde mais dans notre représentation du monde. Elle fournit une explication dans un code de représentation.

Accessibilité de la connaissance : tacite, explicite et implicite

Les auteurs Nonaka et Takeuchi [Nonaka 1995] proposent de diviser la connaissance en deux catégories : la connaissance tacite et la connaissance explicite. La connaissance tacite correspond à la connaissance obtenue à travers l'expérience (les compétences par exemple), à la connaissance simultanée (liée à la situation immédiate) et à la connaissance analogue (aptitude physique). La connaissance explicite correspond à la connaissance rationnelle (dans l'esprit), à la connaissance séquentielle (réaction par rapport à la situation immédiate) et à la connaissance codifiée (production électronique).

Cette catégorisation est étendue dans [Liebowitz 1998] par la proposition d'un troisième niveau de connaissance qui est la connaissance implicite. On accède à la connaissance tacite dans l'esprit humain et dans les organisations à travers un processus d'extraction de connaissance et d'observation de comportements. La connaissance implicite est accessible à partir de

consultations et de discussions. Finalement la connaissance explicite se trouve dans les documents et les systèmes informatiques par l'intermédiaire de formalisation de la connaissance.

Ces différentes typologies de la connaissance permettent d'établir les caractéristiques que celle-ci doit avoir dans le domaine de la RI. Il est primordial qu'elle soit interprétable à la fois par le système et par l'utilisateur du SRI. La connaissance vise à mettre en place un dialogue entre ces deux acteurs. La connaissance doit être active dans le processus de recherche en permettant d'une part au système de sélectionner les granules qu'il restitue et d'autre part à l'utilisateur de situer le contexte et les raisons de cette restitution. De plus, le système peut être enrichi par la connaissance tacite de l'utilisateur comme par exemple ses compétences dans la tâche de recherche qu'il effectue. La spécification de la connaissance implicite présente dans les documents que le SRI manipule peut également l'aider dans la restitution des documents.

Afin que cette connaissance soit intégrée au SRI, elle doit être acquise. La section suivante décrit cette problématique.

3.3 De l'acquisition à l'ingénierie

La connaissance qui peut être fournie à un système informatique a tout d'abord besoin d'être capturée et modélisée. Le domaine de l'Ingénierie des Connaissances (IC) a une finalité applicative reposant sur cette problématique. L'IC est définie dans [Charlet 2000] comme « l'étude des concepts, méthodes et techniques permettant de modéliser et/ou acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches ne se formalisant a priori pas ou peu ».

L'IC a pris la place du domaine de l'Acquisition des Connaissances à partir des années 80. L'évolution des procédés liés à l'acquisition de connaissances peut s'analyser à travers l'évolution de ces deux domaines de recherche.

Acquérir des connaissances est une tâche difficile, dont l'objectif est d'expliciter et de capturer des connaissances explicites sans introduire de biais. Avant la naissance du domaine de l'IC, la connaissance utilisée par les Systèmes Experts était celle d'un expert qui la codait directement dans un langage de représentation. Cette démarche a été remise en cause pour laisser place à de nouveaux systèmes reposant sur une connaissance construite coopérativement avec un ou plusieurs experts à partir d'un modèle de la connaissance. Dans les années 1990, le modèle était un modèle réel du monde tel qu'il était observé à partir des connaissances des experts du domaine ou d'autres sources. Après 1990, un modèle n'a plus uniquement pour objectif de représenter une observation du monde mais une interprétation liée à l'opération que l'on souhaite faire avec la connaissance. Un modèle est alors une abstraction qui permet de réduire la complexité de la modélisation en se focalisant sur certains aspects en fonction du but à atteindre. Le modèle conceptuel est alors défini pour manipuler des objets du monde ainsi que l'interprétation des résultats de la manipulation. Un modèle conceptuel en IC repose sur trois niveaux de connaissance [Shadbolt 1993]. Il exprime tout d'abord comment une tâche va être effectuée. Il utilise également la connaissance d'un domaine qui définit les concepts à manipuler et leurs relations. Finalement, le modèle explicite la manière dont le système résout le problème à partir de la connaissance qu'il utilise.

3.4 Représentation de la connaissance

Le processus d'ingénierie des connaissances définit des étapes pour organiser la connaissances au sein de représentations formelles. Un modèle conceptuel de la connaissance est ensuite traduit en une représentation qui pourra être manipulée par les systèmes informatiques.

Représenter la connaissance a pour objectif de modéliser la connaissance en omettant certains détails non significatifs pour en permettre une meilleure manipulation [Kayser 1997]. Cette question est au cœur des travaux en Intelligence Artificielle. La représentation ne correspond pas à l'entité dans son intégralité. Prenons par exemple une carte routière, son intérêt est de représenter une région afin de pouvoir prévoir un déplacement. Une carte à taille réelle n'aurait aucun intérêt.

Une représentation est une structure composée de symboles construite à partir d'un ensemble de règles de formation [Kayser 1997]. L'ensemble des règles de formation est défini par le langage de représentation choisi. Un ordinateur gère des symboles, il est médiateur de la connaissance, comme l'est un livre. L'utilisateur de l'ordinateur accède, lui, à la sémantique associée à la représentation [Bachimont 1999]. Pour la RI, cette représentation doit intégrer les termes permettant de détecter la connaissance dans les documents.

La représentation des connaissances utilisée dans les Systèmes Experts reposait sur des règles logiques. Le domaine de l'IC a dépassé la problématique des Systèmes Experts pour proposer de nouveaux formalismes pouvant représenter la richesse sémantique de la connaissance en amont de sa représentation formelle opérationnelle. La représentation de la connaissance s'appuie alors sur des représentations au niveau conceptuel pouvant modéliser la « structure cognitive » d'un domaine [Guarino 1994]. Par niveau conceptuel, on entend ici une formalisation sur la description des connaissances avant de se préoccuper de la manière dont un système inférentiel pourra les traiter. Les ontologies définies dans la section 4 sont des exemples de telles représentations. Les langages à base de Frame [Minsky 1975], les logiques de description [Brachman 1985] et les graphes conceptuels [Sowa 1984] sont des langages permettant ces représentations. Ces langages seront décrits dans la section 4.3. Ils ont en commun de donner priorité au pouvoir d'expression par rapport à la capacité de raisonnement logique. Ils permettent de représenter pour un domaine de connaissance donné, les concepts, les relations entre les concepts, ainsi que la sémantique de ces relations. Ces dernières notions sont explicitées dans la section 4.2.1.

4 Représentation de la connaissance et ontologie

Gruber [Gruber 1993] introduit la notion d'ontologie comme "une spécification explicite d'une conceptualisation". Cette définition a été légèrement modifiée par Borst [Borst 1997]. Une combinaison des deux définitions peut être résumée ainsi : « une spécification explicite et formelle d'une conceptualisation partagée ». Cette définition s'explique ainsi [Studer 1998] : *explicite* signifie que le « type des concepts et les contraintes sur leurs utilisations sont explicitement définies », *formelle* se réfère au fait que la spécification doit être lisible par une machine, *partagée* se rapporte à la notion selon laquelle une ontologie « capture la connaissance consensuelle, qui n'est pas propre à un individu mais validée par un groupe », *conceptualisation* se réfère à « un modèle abstrait d'un certain phénomène du monde reposant sur l'identification des concepts pertinents de ce phénomène ».

Une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine [Aussenac 2004].

Les ontologies sont utilisées dans de nombreux domaines. Les domaines recensés en 1998 par Guarino [Guarino 1998b] sont l'ingénierie des connaissances, la modélisation qualitative, l'ingénierie des langages, la conception de bases de données, la recherche d'information, l'extraction d'information, la gestion et l'organisation de connaissances. Depuis, grâce à l'essor du Web, elles sont utilisées dans le domaine de l'e-commerce et sont au cœur du Web Sémantique [Berners-Lee 2001], future version du Web actuel. Un des plus grands projets reposant sur

l'utilisation des ontologies consiste à ajouter au Web une véritable couche de connaissance permettant des recherches d'information au niveau sémantique et non plus au simple niveau lexical et/ou syntaxique. A terme, il est prévu que des applications déployées sur l'Internet pourront mener des raisonnements utilisant les connaissances stockées sur la toile.

Derrière l'utilisation d'ontologies dans ces différents domaines, se cachent en fin de compte plusieurs représentations de connaissances. Ces représentations peuvent être distinguées suivant deux axes : la nature de la connaissance représentée dans l'ontologie et le degré d'engagement sémantique qui a motivé la formalisation de l'ontologie. Le premier axe fait en particulier référence au type de connaissances représentées (génériques, de domaines ou liées à la tâche). Le second axe fait en particulier référence au niveau sémantique des connaissances que l'ontologie représente (ressource terminologique versus ressource conceptuelle). Nous présentons ces deux aspects : nature des connaissances et engagement sémantique dans les sections suivantes.

4.1 Nature des connaissances

La première distinction à faire sur les représentations de connaissance associées à la notion d'ontologie repose sur la nature des connaissances représentées dans l'ontologie. La nature de ces connaissances peut varier soit par rapport à la structure de la connaissance (section 4.1.1.) soit par rapport au contenu de la connaissance (section 4.1.2).

4.1.1 Différentes structures de la connaissance

La connaissance contenue dans l'ontologie peut représenter plusieurs structures. La classification décrite dans [Heijst 1997] distingue trois types d'ontologies suivant ce critère.

- Les ontologies terminologiques ou linguistiques spécifient les termes utilisés pour représenter la connaissance d'un domaine. Un exemple de ce type d'ontologie est le réseau sémantique UMLS (Unified Medical Language System) [Lindeberg 1993].
- Les ontologies de l'information spécifient la structure des enregistrements d'une base de données. Les schémas de base de données en sont un exemple. Elles proposent un cadre de représentation de la connaissance stockée mais ne spécifient pas de détails sur la sémantique des champs.
- Les ontologies pour la modélisation de la connaissance spécifient la conceptualisation de la connaissance. Ces ontologies ont une structure beaucoup plus riche que celle des deux autres types. Elles sont généralement conçues en fonction de l'utilisation prévue de la connaissance qu'elles contiennent.

Pour la RI, la structure de connaissances utile se situe entre celle des ontologies terminologiques et celle des ontologies pour la modélisation de la connaissance. Elles ont pour but de définir les termes liés à la connaissance pour que celle-ci soit décelable dans les documents. Mais elles doivent également permettre d'interpréter la connaissance à partir d'un niveau conceptuel afin que des mécanismes élaborés puissent être intégrés au SRI.

4.1.2 Différents contenus

Un autre critère pour la classification des ontologies est le contenu de la connaissance qu'elles représentent, c'est-à-dire le sujet de la conceptualisation [van Heist 1997], [Guarino 1998b].

- Les **ontologies génériques** définissent des concepts considérés comme génériques à plusieurs domaines. WordNet [Miller 1988] par exemple est une ontologie dont le but est de représenter la langue naturelle anglaise. WordNet est un système de références lexicales dont la conception a été inspirée par les théories de la mémoire linguistique humaine. Elle est composée d'ensembles de synonymes appelés *synsets*, où chaque terme est regroupé en classes d'équivalence sémantique. Chaque ensemble de synonymes représente un concept particulier. Chaque terme appartient de plus à une catégorie lexicale donnée (nom, verbe, adverbe, adjectif). Un terme peut appartenir à plusieurs synsets et à plusieurs catégories lexicales. Les ensembles de synonymes sont associés par des relations sémantiques : généricité/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.

  Par exemple, le figure 1.2 présente l'ensemble des différents sens retrouvés pour le mot **dispersion**

> 1.**dispersion**, scattering -- (spreading widely or driving off)
>
> 2.distribution, **dispersion** -- (the spatial property of being scattered about over an area or volume)
>
> 3.**dispersion**, dispersal, dissemination, diffusion -- (the act of dispersing or diffusing something; "the dispersion of the troops"; "the diffusion of knowledge")
>
> Les sens répertoriés renvoient au sens dans le langage courant (sens 1 et 3) ainsi qu'au sens du mot dans le domaine scientifique, plus précisément le domaine de la physique (sens 2).

Figure 1.2 : ensemble des différents sens du mot dispersion dans WordNet.

  Nous partageons le point de vue de Charlet [Charlet 2002] suivant lequel la limite de ces ontologies générales est leur difficile réutilisation car elles ont pour objectif de recouvrir tous les sens des mots et ne normalisent pas leur sens.

  La normalisation sémantique consiste à organiser au sein d'un modèle conceptuel des connaissances, à partir de la compréhension du domaine et de l'application visée. Cela revient à associer aux termes une signification qui fait abstraction des variations de sens liées à d'autres domaines. Cette abstraction du contexte conduit à construire des concepts, considérés en tant que «signifiés non contextuels», normés au sens où ils sont décrits selon un certain point de vue (celui de la tâche, qui fixe un contexte de référence).

- Les **ontologies de domaine** sont des conceptualisations spécifiques à un domaine particulier. Les méthodes actuelles d'acquisition de la connaissance font la distinction explicite entre connaissance du domaine et ontologie du domaine. La connaissance du domaine décrit des situations factuelles du domaine alors que l'ontologie pose des contraintes sur la structure et le contenu de la connaissance du domaine. Comparées aux ontologies génériques, les ontologies de domaine ont pour avantage de permettre une normalisation des concepts dans le cadre du domaine considéré et donc, selon nous, de permettre une meilleure représentation de la connaissance. L'ontologie Ménélas [Zweigenbaum 1993] est un exemple d'ontologie de domaine, celui des maladies coronariennes, rassemblant des concepts et leurs relations structurés à partir de la relation « sorte de ». Ménélas comprend également des lexiques sémantiques et morpho-syntaxiques des mots simples et composés. Cette ontologie est dédiée à l'analyse automatique de comptes-rendus d'hospitalisation.

- Les **ontologies d'application** contiennent toutes les définitions qui sont nécessaires pour modéliser la connaissance propre à l'élaboration d'une tâche particulière. Généralement, les ontologies d'application combinent des éléments d'ontologies de domaine et d'ontologies génériques choisies en fonction des méthodes spécifiques pour réaliser la tâche visée. Elles sont rarement réutilisables pour une autre application.
- Les **ontologies de représentation de la connaissance** permettent d'expliquer la conceptualisation sous-jacente aux formalismes de représentation [Davis 1993]. Elles proposent un cadre de représentation sans émettre d'hypothèse sur le monde. On les désigne également comme ontologies abstraites ou de haut niveau parce qu'elles permettent de définir des concepts abstraits et peuvent être re-utilisées pour définir des concepts spécifiques. Un exemple d'ontologie de ce type est la Frame Ontology utilisée dans Ontolingua [Gruber 1993].

L'objectif de nos travaux étant d'intégrer dans un SRI des ontologies liées à un domaine de connaissance donné, les ontologies que nous considérons dans la suite du mémoire sont des ontologies de domaine. Ces ontologies permettent de normaliser la connaissance manipulée par le système par rapport à la connaissance qui lui est utile.

4.2 Engagement sémantique

La deuxième distinction à faire sur les représentations de connaissance induites par la notion d'ontologie repose sur le degré d'engagement sémantique de celles-ci. Le degré d'engagement sémantique correspond au niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique [Bachimont 2000]. La figure 1.3 présente un spectre des différentes représentations en fonction de ce degré. L'ensemble de celles-ci porte sur un domaine de connaissance. A la gauche du spectre, les représentations correspondent à des représentations dont la sémantique est uniquement définie dans l'esprit des personnes qui les utilisent. A la droite du spectre, les représentations définissent une sémantique formelle et explicite interprétable par l'homme et la machine.

Afin d'expliciter ce niveau d'engagement puis de décrire chacune de ces représentations, il est important de les distinguer en définissant plusieurs notions impliquées dans leur degré de formalisation. Celles-ci sont décrites dans la section 4.2.1. Les différentes représentations seront décrites dans la section 4.2.2. Elles sont distinguées suivant deux principes : les ressources faisant intervenir des termes et les ressources composées de concepts.

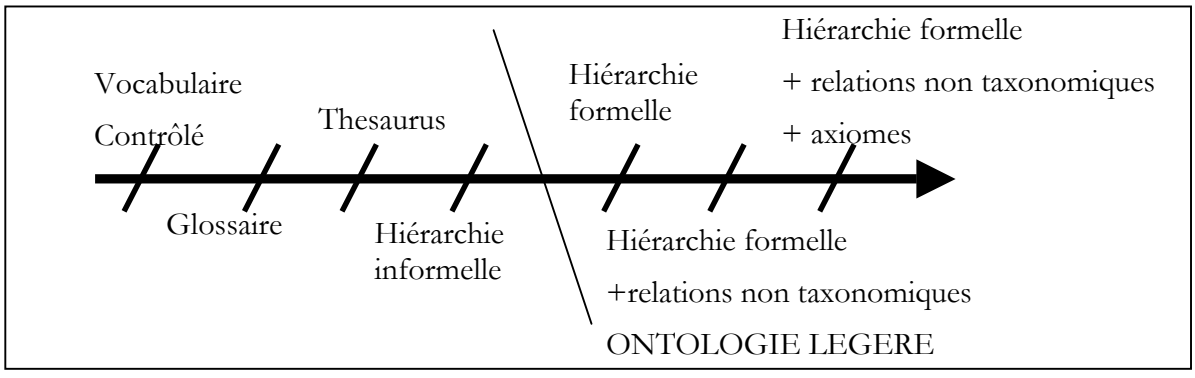


Figure 1.3 : Les différentes représentations de connaissance à partir de leur degré d'engagement sémantique inspiré de [Lassila 2001]

4.2.1 Notions sous-jacentes

Afin d'expliciter les engagements sémantiques dans la formalisation d'ontologies, il convient de préciser les notions de concept, relation, subsomption et axiome.

4.2.1.1 Concept

Un concept se définit par Bachimont à trois niveaux [Bachimont 2004]. Un concept est une signification. Sa place dans un système de significations permet de le comprendre, de le distinguer et de le différencier par rapport à d'autres concepts. Un concept est une construction. Comprendre un concept revient à construire l'objet dont il est le concept. Un concept est une prescription. On le comprend en exécutant l'action qu'il entreprend.

Uschold partage ce point de vue et définit de façon plus pragmatique la notion de concept. Un concept représente pour un objet matériel, une notion ou une idée [Uschold 1995]. Il est composé de trois parties : un ou plusieurs **termes**, une **notion** et un **ensemble d'objets**. La notion correspond à la sémantique du concept, elle est définie à travers ses propriétés et ses attributs. La notion est appelée **intention** du concept. L'ensemble d'objets correspond aux objets définis par le concept, il est appelé **extension** du concept ; les objets sont les **instances** du concept. Le ou les **termes** permettent de désigner le concept. Ces termes sont aussi appelés **labels** de concept. Par exemple, le terme « lapin » renvoie à un animal possédant de longues oreilles et une queue et à l'ensemble des objets ayant cette description. Afin que les concepts soient reconnus de façon non ambiguë par la machine, il est souhaitable qu'un concept soit identifié à partir de plusieurs termes, ce qui permet de gérer la synonymie et de les désambiguïser les uns par rapport aux autres [Gomez-Perez 1996].

Un concept est défini à partir d'une sémantique référentielle (due à son extension) et une sémantique différentielle (due à son intention). Un concept peut avoir une extension vide, c'est le cas des concepts génériques ou abstraits comme par exemple « la vérité ». Deux concepts peuvent avoir la même extension et des intentions différentes. C'est le cas par exemple des « lapins » considérés comme « animaux de compagnie » ou bien comme « ressource culinaire ». Il est considéré par certains auteurs que l'intention d'un concept permet à elle seule de définir le sens d'un concept [Guarino 1994] ; d'autres auteurs considèrent que le sens dépend de l'intention et de l'extension du concept [Kassel 1999].

4.2.1.2 Relation sémantique

Une relation sémantique R représente un type d'interaction entre les concepts d'un domaine c_1, c_2, .., c_n. Elle se définit formellement à partir d'un produit de n concepts : R : $c_1 \times c_2 \times \ldots \times c_n$; « subsume », « est un phénomène lié à » sont des exemples de relations binaires.

Les relations les plus courantes dans la littérature sont les relations d'équivalence, taxonomiques, patronymiques, de dépendance, topologique, causale, fonctionnelle, chronologique [Gomes-Peres 2000].

Relation taxonomique (ou subsomption)

La notion de subsomption (aussi appelée relation « est un », relation taxonomique ou relation de spécificité/généricité) est une relation binaire particulière qui implique l'engagement sémantique suivant [Guarino 2001] : un concept c1 subsume un concept c2 si toute relation sémantique de c1 est aussi relation sémantique de c2, en d'autres termes si le concept c2 est plus spécifique que le concept c1. Les instances se rapportant au concept c2 seront des instances de c1, par contre une partie seulement des instances de c1 seront des instances de c2. La notion abordée par le concept c2 (intention du concept) sera plus précise que celle abordée par c1. La relation de subsomption permet d'organiser hiérarchiquement un ensemble de concepts.

La relation de subsomption est une **relation d'ordre partiel** définie à partir des propriétés suivantes:

- L'**asymétrie** : cette propriété signifie que l'inclusion d'une classe d'individus X dans une classe d'individus Y implique que Y n'est pas incluse dans X.

  Formellement, cette propriété garantit que : X subsume Y, si et seulement si non (Y subsume X),

- La **transitivité** : soit une classe d'individus X qui subsume une classe Y, qui elle-même subsume Z, alors X subsume Z.

  Formellement : (X subsume Y) et (Y subsume Z) $\Rightarrow$ (X subsume Z)

- La **non réflexivité** : Cette propriété implique qu'un fait décrit par la relation « est un » ne peut pas s'écrire de plusieurs façons.

  Formellement : non (X subsume X)

L'**héritage multiple** est une propriété qui peut être définie sur la relation de subsomption : un concept d'une ontologie peut avoir plusieurs pères par la relation de subsomption. L'héritage multiple implique que le concept hérite des propriétés de tous ses pères.

Relation associative

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

4.2.1.3 Axiome

Les axiomes ont pour but de définir dans un langage logique la description des concepts et des relations permettant de représenter leur sémantique. Ils représentent les intentions des concepts et des relations du domaine et, de manière générale, les connaissances n'ayant pas un caractère strictement terminologique [Staab 2000]. Les axiomes sont des expressions qui sont toujours vraies. Leur inclusion dans une ontologie peut avoir plusieurs objectifs : définir la signification des composants, définir des restrictions sur la valeur des attributs, définir les arguments d'une relation, vérifier la validité des informations spécifiées ou en déduire de nouvelles.

La figure 1.4 présente des exemples d'axiomes formalisés à partir du langage OWL-Lite. Ce langage est décrit dans la section 4.3.

```
  <owl:Class rdf:ID="TexasThings">
 <owl:equivalentClass>
  <owl:Restriction>
   <owl:onProperty rdf:resource="#locatedIn" />
   <owl:someValuesFrom rdf:resource="#TexasRegion" />
  </owl:Restriction>
 </owl:equivalentClass>
</owl:Class>
```

Dans ce premier exemple, le concept Texas Things est restreint aux individus localisés dans la région Texas par l'axiome en gras.

```
<owl:AllDifferent>
 <owl:distinctMembers rdf:parseType="Collection">
  <vin:WineColor rdf:about="#Red" />
  <vin:WineColor rdf:about="#White" />
  <vin:WineColor rdf:about="#Rose" />
 </owl:distinctMembers>
</owl:AllDifferent>
```

Dans cet exemple, l'axiome permet de définir une classe composée des différentes couleurs de vin.

Figure 1.4 : Exemples d'axiomes formalisés à partir de OWL-Lite tirés de http://www.w3.org/TR/owl-guide/

4.2.2 Ressources terminologiques

Les ressources terminologiques ne définissent pas des concepts mais des termes. Contrairement à leur utilisation dans la définition des concepts, dans ces ressources, les termes ne font pas référence à des notions (intention du concept) et des objets (extension du concept) mais définissent uniquement le vocabulaire lié à la connaissance représentée. La notion de terme est intentionnellement choisie par rapport à la notion de mot. La notion de mot désigne une unité textuelle, alors que la notion de terme fait référence à l'ensemble des variantes lexicales d'un mot ou d'un groupe de mots.

4.2.2.1 Vocabulaire contrôlé

Soit T un ensemble de termes d'un domaine. Un vocabulaire VC contrôlé est défini par $VC=\{t1,t2,..,tn / ti \in T\}$.

Un vocabulaire contrôlé est un ensemble de termes définis par un groupe de personnes ou une communauté. La signification des termes n'est pas forcément définie et il n'y a pas d'organisation logique entre les termes [Lassila 2001]. Ce vocabulaire peut être utilisé afin de labelliser des contenus documentaires. Les catalogues sont des exemples de vocabulaires contrôlés.

4.2.2.2 Glossaires

Soit T un ensemble de termes d'un domaine, soit D un ensemble de définitions en langage naturel. Un glossaire G est défini par le couple (T,D).

Un glossaire est un ensemble de termes avec leur signification. La définition de chaque terme est donnée en langage naturel. Cette représentation apporte plus d'informations car une personne peut lire la définition, cependant elle n'est pas interprétable par l'ordinateur [Lassila 2001].

4.2.2.3 Hiérarchie informelle

Soit T un ensemble de termes et R une relation de T x T où R(t1,t2) signifie que le terme t1 est plus général que le terme t2 ou que le terme t2 est plus spécifique que le terme t1. Une hiérarchie informelle est définie par le couple (T,R).

Les hiérarchies informelles sont des hiérarchies explicites organisant des catégories à partir de la notion générale de généralisation / spécification. Elles ont fait leur apparition sur le Web comme par exemple la hiérarchie proposée par Yahoo[5]. Cependant, ces hiérarchies ne sont pas formelles car la hiérarchisation des catégories ne respecte pas la stricte notion de subsomption (définie dans la section 4.2.1.3) [Lassila 2001]. Tout d'abord les termes employés pour désigner les catégories ne permettent pas de définir clairement le sens de la catégorie. Prenons par exemple l'extrait de la hiérarchie Yahoo suivante :

Accueil > Mode & Accessoires > Pour la Femme > Tous les Accessoires Femme > Pierres / Perles > Perle .

Le terme Perle définissant la catégorie la plus profonde dans cette branche de la hiérarchie ne fait pas référence à la notion de perle dans son ensemble ; elle devrait être désignée par les termes « accessoires féminins contenant des perles ».

De plus, au sens strict de la subsomption, les individus de cette dernière catégorie devraient avoir les mêmes propriétés sémantiques que celles de la classe Pierres/Perles. Ce type de hiérarchies regroupant des catégories disjointes (accessoires en pierre ou en perles) rend problématique l'héritage des propriétés.

4.2.2.4 Thésaurus

Soit T un ensemble de termes et $\mathfrak{R}$ un ensemble de relations de T x T. Un thésaurus est défini par le couple (T, $\mathfrak{R}$).

Un thésaurus est un ensemble de termes organisés suivant un nombre restreint de relations [Foskett 1980]. Les relations entre termes les plus typiques sont présentées dans la figure 1.5 [Foskett 1980], [Miles 2005], [Soergel 2004]. Elles définissent des relations entre termes synonymes (terme préféré, terme à utiliser à la place de), entre termes préférés (terme plus spécifique, terme plus générique, terme lié à). Afin d'uniformiser leur format de représentation, différentes normes spécifient les thésaurus monolingues (ISO 2788:1986[6], AFNOR NF Z47-100:1981 , ANSI Z39[7]) et multilingues (AFNOR NF Z47-101 :1990, ISO 5964 :1985[8]).

---

[5] www.yahoo.fr

[6] http://www.collectionscanada.ca/iso/tc46sc9/standard/2788e.htm

[7] http://www.niso.org/standards/resources/Z39-19.html

[8] http://www.collectionscanada.ca/iso/tc46sc9/standard/5964e.htm

t_1 **Terme préféré dans** $t_2, t_3,\ldots,t_N$	t_1 est le terme préféré pour désigner l'ensemble des synonymes $t_2, t_3,\ldots,t_N$.
t_1 **Note** texte	Remarque sur le terme t_1(usage exceptionnel, contexte d'utilisation)
t_1 **Utiliser plutôt** t_2	t_2 est utilisé pour désigner t_1
t_1 **Utilisé pour** t_2	t_1 est utilisé pour désigner t_2
t_1 **Plus spécifique que** t_2	Le terme désigné par t_1 est plus spécifique que le terme désigné par t_2
t_1 **plus générique que** t_2	Le terme désigné par t_1 est plus générique que le terme désigné par t_2
t_1 **est lié à** t_2	t_1 est un terme lié ou associé à t_2

Figure 1.5 : les relations entre termes les plus typiques dans un thésaurus

Les thésaurus sont principalement utilisés pour assister les documentalistes dans la tâche d'indexation manuelle de documents. Ils sont reconnus pour présenter différents avantages dans ce contexte [Foskett 1977]. Ils offrent tout d'abord une vue générale sur les termes et relations d'un domaine. Ils définissent ensuite un vocabulaire standardisé pour l'indexation. Ils permettent d'assurer qu'un seul terme d'un ensemble de synonymes soit choisi pour l'indexation (terme dit « à utiliser »). Ils sont également utilisés lors de la spécification d'une requête pour spécifier ou généraliser une recherche documentaire à partir des termes dits plus spécifiques ou plus génériques.

De nombreux auteurs [Tudhope 2001], [Soergel 1974], [Fischer 1998] ainsi que la description du standard ISO 2788 considèrent que les relations de généricité /spécificité définissant la hiérarchie des thésaurus ne suivent pas l'engagement sémantique impliqué par la relation de subsomption. Ils considèrent que les relations *termes spécifiques*, *termes plus génériques* regroupent différentes relations sémantiques telles que les relations de généricité mais aussi « partie de » et « instance de ». Fischer explique cette ambiguïté par le fait que la définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation, recherche), et non par la formalisation de la connaissance du domaine [Fischer 1998]. Il prend comme référence la définition donnée dans [Soergel 1974] : « Le terme A est considéré comme étant plus générique que le terme B si pour toute recherche inclusive sur le terme A tous les éléments traitant de B doivent être retrouvés. Inversement B est plus spécifique ». La définition introduit donc de la subjectivité et implique un jugement de l'expert sur le résultat d'une recherche. Les thésaurus sont depuis de nombreuses années utilisés en RI [Baeza-Yates1999]. Cependant, par le manque de formalisation et l'objectif de leur conception (l'aide aux documentalistes), ils présentent un degré d'ambiguïté et les termes qui les composent doivent être interprétés par une personne pour pouvoir capturer la sémantique implicite qu'ils sous-entendent.

Les thésaurus à facettes sont une catégorie de thésaurus. Les termes sont alors organisés suivant plusieurs hiérarchies mutuellement exclusives représentant chacune une facette du domaine représenté. Les termes appartiennent donc à une seule facette mais des relations non hiérarchiques peuvent exister entre les différents termes des différentes facettes [Spiteri 1999]. Le thésaurus AAT[9] est un exemple de ce type de thésaurus. Il représente le domaine de l'architecture

[9] http://www.getty.edu/research/conducting_research/vocabularies/aat/

organisée suivant sept facettes : les concepts associés, les attributs physiques, les styles et périodes, les agents, les activités, les matériaux et les objets.

4.2.3 Ressources conceptuelles

Les ressources conceptuelles témoignent d'un engagement sémantique qui repose sur la notion de concepts définie dans la section 4.2.1.1. Différents niveaux sémiotiques sont à prendre en considération dans une ressource conceptuelle [Maedche 2002]. Le niveau lexical couvre tous les termes ou labels définis pour désigner les concepts. Le niveau conceptuel représente les concepts et la sémantique qui leur est associée à partir des relations conceptuelles entre eux. Une ressource conceptuelle est définie à partir d'une structure qui décrit son niveau conceptuel et d'un lexique correspondant au niveau lexical. L'objectif de nos travaux étant de manipuler et de construire de telles ressources, nous les définissons à partir de notions formelles afin de pouvoir les désigner par la suite.

4.2.3.1 Hiérarchie de concepts

Une structure d'une hiérarchie de concepts est le couple $S_H\{C, \leq^C\}$

Où :

C est un ensemble de concepts,

$\leq^C$: 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)

Le lexique d'une hiérarchie de concepts est le couple L_H : $\{L^C, F\}$

F est une fonction appelée référence tel que $F \rightarrow L^c$ pour les concepts,

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

Une hiérarchie de concepts est le couple (S_H, L_H).

4.2.3.2 Ontologie dites « légères »

La structure et le lexique définissant une ontologie légère sont les suivants :

La structure est un tuple $S_O := \{C, R, A, T, \leq^C, \sigma_R, \sigma_A,\}$ où :

- C, R, A, T sont des ensembles disjoints contenant les concepts, les relations associatives, les relations d'attribut, les types de données
- $\leq^C$: 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)
- σ_R : $R \rightarrow$ C x C est la signature d'une relation associative
- σ_A : $A \rightarrow$ C x T est la signature d'une relation d'attribut

Le lexique est un tuple L_O : $\{L^C, L^R, F, G\}$

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

  $F \rightarrow L^c$ pour les concepts, $G \rightarrow L^r$ 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, relations et instances désignés par un terme et réciproquement.

Une ontologie légère est le couple $O=(S_O, L_O)$. Elle est dite « lègère » car l'engagement sémantique qu'elle suit n'est pas totalement formel dans la mesure où aucun axiome n'est spécifié. Cependant de telles ontologies présentent différents avantages. Elles sont facilement interprétables par l'homme. Leur construction, leur vérification et leur mise à jour demandent moins d'effort. Enfin, il est plus facile de trouver un consensus lors de leur spécification [Kiryakov 2004]. Contrairement aux ontologies lourdes, les ontologies légères ne possèdent pas d'axiomes.

4.2.3.3 Ontologies lourdes

La structure d'une ontologie lourde est un tuple $S_O := \{C, R, A, T, A_x, P_{log}, \leq^C, \sigma_R, \sigma_A\}$.

Ce tuple est défini à partir de la structure d'une ontologie légère à laquelle est ajouté un ensemble d'axiomes A_x. Les axiomes sont décrits à partir des primitives logiques P_{log} définies par le langage logique considéré.

Une ontologie lourde est définie par le couple (S_O, L_O)

TOVE [Gruninger 1995b] et PIF [Lee 1995] sont des exemples d'ontologies rigoureusement formelles. Leur avantage repose sur la réduction considérable des interprétations possibles des concepts et donc la minimalisation des ambiguïtés. Cependant, elles demandent de lourds efforts de conception [Uschold 2003] et ne peuvent couvrir que des domaines précisément définis.

4.2.3.4 Modèle d'un domaine

Afin de compléter la connaissance liée à un domaine, le modèle d'un domaine est représenté à partir d'une ontologie légère ou lourde O telle que présentée dans les sections précédentes et des instances qui sont associées à ses concepts.

Il se formalise par le tuple suivant $(O, I, V, f_C, f_T, f_R, f_A)$ avec :

- O l'ontologie de domaine considéré
- I l'ensemble des instances
- V l'ensemble des valeurs des types de données

- $f_C : I \rightarrow C$ est la fonction d'instanciation d'un concept
- $f_T : V \rightarrow T$ est la fonction d'instanciation d'un type de donnée
- $f_R : I \times I \rightarrow R$ est la fonction d'instanciation d'une relation associative
- $f_A : I \times V \rightarrow A$ est la fonction d'instanciation d'une relation d'attribut

4.3 Langages de représentation des ontologies conceptuelles

L'objectif de nos travaux étant la conception et la prise en compte par le SRI de ressources conceptuelles, nous avons analysé les différents langages de représentation des ontologies conceptuelles. Les langages dédiés aux ontologies sont principalement issus des formalismes liés aux réseaux sémantiques. Nous les décrivons dans la section suivante. Nous nous concentrons ensuite sur les langages de représentation qui en sont issus.

4.3.1 Réseaux sémantiques et langages associés

Un **réseau sémantique** est une représentation graphique d'une conceptualisation d'une (ou plusieurs) connaissance humaine [Quillian 1968]. Il est représenté sous la forme d'un graphe étiqueté et orienté. Un arc lie un nœud de départ à un nœud d'arrivée. Chaque nœud peut être relié par un ou plusieurs arcs. Les inférences possibles dépendent de la nature des liens. Cependant, ce type de définition ne concerne que la structure du graphe et ne permet pas d'ajouter de l'information sémantique. De nombreuses études [Woods 1975], [Brachman 1977] ont montré que ce type de graphe manque de précision sémantique et mène à des confusions entre les relations et aussi entre les classes et individus. Elles ont mené à la définition de nouveaux formalismes tels que les frames, les logiques de description et les graphes conceptuels.

Les **frames** [Minsky 1975] sont présentés comme étant une structure de données capable de représenter des objets structurés. Un frame représente donc une classe ou un objet. Les frames sont organisés dans une hiérarchie suivant un lien de spécification. Les composants du frame sont appelés « slots », ils sont considérés comme des attributs de la structure. Ils peuvent être de plusieurs natures : valeur de l'attribut (qui peut être vide), ensemble de valeurs, restriction de valeurs, valeur par défaut, une propriété avec un autre frame, une combinaison des différents cas. L'intérêt des frames est qu'ils permettent de représenter la façon de penser d'experts en fournissant une représentation structurée et concise des relations utiles [Fikes 1985]. L'information peut être partagée entre plusieurs frames grâce à l'héritage.

Les **logiques de description** issues des frames reposent sur trois notions de base : les concepts représentant des classes (ensemble d'objets), les rôles (relations liant deux objets) et les individus (objets représentant les classes qu'ils instancient). Pour décrire ces éléments, deux structures sont utilisées : la *T-BOX* et la *A-BOX*. La *T-BOX* (boîte terminologique) comprend la description des concepts et des rôles. Cette description est structurée à l'aide du lien hiérarchique *sorteDe*. Deux concepts particuliers figurent au minimum dans la *T-BOX* : le concept le plus générique (anything) et le concept le plus spécifique (nothing). La *A-BOX* (boîte assertionnelle) est constituée des individus, de leur description et des règles qui leur sont attachés. Les inférences reposent sur la reconnaissance d'instances de concepts à partir de leur définition, la détection des concepts plus généraux ou plus spécifiques, et la classification ordonnant les concepts dans la hiérarchie. Les logiques de description sont plus flexible que les frames et reposent sur une sémantique et une syntaxe rigoureuses [Baader 1991]. Elle est utilisable en RI car elle permet de traiter des données erronées ou incomplètes tout en offrant la possibilité d'ordonner hiérarchiquement les données [Todirasscu 2001]. Cependant, elles nécessitent l'élaboration manuelle de ressources de connaissances formalisées à partir de cette logique.

Les **graphes conceptuels** ont été présentés par Sowa en 1984 [Sowa,1984] et utilisent une notation à base de graphes. Ils ont été définis comme un langage pivot entre le langage

naturel et la logique du premier ordre. Ils visent à formaliser les relations entre prédicats et arguments dans une phrase. Ils sont composés de deux types de nœuds étiquetés : les nœuds concepts et les nœuds relations. Les nœuds concepts et les nœuds relations sont respectivement typés par des types de nœuds et des types de relations, organisés suivant un ordre partiel. Les graphes conceptuels peuvent être vus comme des schémas permettant de représenter graphiquement des formules logiques, ou bien des schémas sans contraintes, servant juste d'interface « graphique » à la représentation de formules ou bien comme des graphes munis d'opérations de graphes permettant le raisonnement et leur manipulation en s'appuyant sur la théorie des graphes. Les graphes conceptuels ont été utilisés dans les systèmes d'information pour la représentation de requêtes et de documents dans [Guarino 1999]. Ils sont élaborés manuellement et une ressource lexicale (Wordnet) est utilisée pour les mettre en correspondance avec les requêtes de l'utilisateur.

Les ressources conceptuelles que nous souhaitons manipuler dans nos travaux doivent permettre un traitement efficace des collections volumineuses traitées dans le domaine de la RI. Nous souhaitons donc les utiliser comme indexation automatique des documents. Les formalismes que nous venons de présenter sont difficilement adaptables dans ce cas-là.

4.3.2 Langages de représentation d'ontologie

Différents langages de spécification d'ontologies issus des formalismes précédemment présentés sont apparus à partir des années 1990, tels que CycL et KIF [Genesereth 1994], LOOM [MacGregor 1991], F-Logic [Kifer 1995] et OCML [Shabot 1993]. Pour une évaluation détaillée des langages précédemment présentés voir [Su 02]. Nous nous concentrons dans cette section sur les langages orientés Web Sémantique. La raison de ce choix est que ces langages sont ou ont été pour la plupart recommandés par le W3C[10].

- **XML** [Bradley 2001] est un langage permettant de générer des balises pour la structuration de données et de documents. Il permet la représentation et l'échange de documents semi-structurés. **XML-schéma** [Fallside 2001] permet de définir la structure, les contraintes, et la sémantique de documents XML. Ce langage n'est pas vu comme un langage d'ontologies car il a été créé pour vérifier la structure de documents XML. Les primitives qu'il met en place sont plutôt orientées application que concept. En effet, la sémantique définie dans le document est interprétable dans le contexte de l'opération faite sur le document mais ne permet pas d'établir des inférences en dehors de ce contexte. XML et XML-schéma sont considérés comme des langages définissant le format de « message » alors qu'un langage d'ontologies a pour but de « représenter » la connaissance.
- **RDF** [Lassila 1999] permet d'encoder, d'échanger et de réutiliser des méta-données structurées. Il a été créé pour gérer les méta-données de documents XML mais peut également être utilisé pour des ontologies. Il permet de définir des ressources avec des propriétés et des états. **RDF-Schéma** définit les relations entre ces ressources. Le pouvoir sémantique de ces deux langages est limité car les axiomes ne peuvent pas être directement décrits. Le type des relations (symétrique, transitive, …) ne peut être spécifié.
- **OIL** (Ontology Inference Layer) [Decher 2000] est à la fois un langage de représentation et d'échange pour les ontologies. Il combine les primitives des langages reposant sur les frames avec une sémantique formelle et des possibilités de raisonnement issues de la logique de description. Pour être utilisé sur le Web, il repose sur les standards RDF(S) et

[10] http://www.w3.org/

XML. La description de l'ontologie est divisée en trois couches : la couche objet (instances concrètes), la couche de premier méta-niveau (définition de l'ontologie) et la couche de second meta-niveau (définition des caractéristiques de l'ontologie). OIL permet de définir des classes et des relations et un nombre limité d'axiomes. Les relations sont considérées comme des classes et peuvent être organisées hiérarchiquement.

- **XOL** (XML based Ontology Exchange Language) [Karp 1999] a été créé pour échanger des ontologies se rapportant à la biologie moléculaire mais est applicable à d'autres domaines. Cependant, les relations entre concepts ne peuvent pas être spécifiées correctement.
- **SHOE** (Simple HTML Ontology Extensions) [Luke 00] est une extension de HTML qui permet de rajouter de la sémantique dans ce type de documents. Il permet de définir des primitives pour spécifier et étendre les ontologies et annoter les documents Web. Chaque page déclare quelle ontologie elle utilise. L'inconvénient de ce langage est que les annotations des documents sont stockées à leur niveau et ne peuvent être centralisées.
- **DAML+OIL** [Horrocks 2001] a été proposé par le W3C pour représenter des méta-données et des ontologies. DAML a été transformé en DAML+OIL en intégrant certaines propriétés de OIL [Decher 2000]. Il repose sur RDF et RDF schéma et fournit en plus des primitives plus riches issues de la logique de description. Les frames définis dans OIL ont été pour la plupart supprimées et remplacées par les assertions faites à l'aide d'un ensemble limité d'axiomes. Le résultat est que le langage est mieux adapté que RDF à l'utilisation et la maintenance d'ontologies mais présente des limites quant à la construction d'ontologie [Bechhofer & al 01].
- **TOPIC MAPS**[11] ont été créés par la Convention for Application of HyTime (CapH) dont le but était de développer une application automatique d'indexation de livres. Les TM ont été acceptés par le groupe SGML d'ISO en 1996 et standardisés en janvier 2000 [ISO 13250]. Topic Maps est un standard permettant de formaliser la sémantique sous la forme de méta-données. Il est défini à partir de thématiques (topics), d'occurrences de ces thématiques et d'associations non directionnelles entre les thématiques. Le rôle de chacun des membres de l'association a donc besoin d'être spécifié. Un mécanisme propre aux Topics Maps permet de préciser le contexte dans lequel l'association est valable ou intéressante. RDF et Topics Maps ont été créés dans le même but : décrire et organiser des méta-données. Ils sont compatibles et « traduisibles » de l'une à l'autre forme. Cependant, les TM présentent de nombreux intérêts : le mécanisme de spécification du contexte n'existe pas dans RDF, les TM permettent de connaître la relation et le rôle de l'objet dans la relation alors qu'avec RDF il est difficile de savoir si la source est un concept en relation avec l'objet ou contenant de l'information sur l'objet.
- **OWL Ontologie Web Language**[12] [McGuinness 2004] est le standard actuellement proposé par le W3C pour représenter les ontologies. Il a été créé pour être utilisé par les applications cherchant à traiter le contenu de l'information et non plus uniquement à présenter l'information. OWL se veut plus représentatif du contenu du Web que XML, RDF et RDF-Schéma en apportant un nouveau vocabulaire avec une sémantique formelle. OWL est une révision de DAML+OIL définie d'après l'expérience acquise lors de la création et l'utilisation de ce langage. OWL ajoute du vocabulaire pour décrire les propriétés et classes, comme par exemple la disjonction de classe, la cardinalité (exactement un), l'égalité, les types de propriétés plus riches, les caractéristiques de

[11] http://www.topicmaps.org/

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

propriété (symétrie, transitivité, …) et les classes énumérées. OWL est décliné en trois sous langages d'expressivité croissante : OWL lite, OWL DL, OWL Full. OWL Lite est fait pour des besoins préliminaires permettant de définir une hiérarchie et des contraintes simples. Il permet de définir facilement des thésaurus ou taxonomies. OWL DL et Full reposent sur OWL Lite auquel sont ajoutés des constructeurs supplémentaires. OWL DL supporte des besoins d'expressivité maximaux tout en garantissant une complétude de calculs et de décidabilité nécessaires aux systèmes de raisonnement. Il repose sur les éléments OWL auxquels il associe un grand nombre de restrictions (par exemple, une classe peut être une sous-classe de nombreuses autres classes, mais pas une instance d'une classe). OWL DL est conçu pour pouvoir supporter la logique de description. Cette logique appartient à un domaine de recherche qui a pour but d'aider au raisonnement sur une base de connaissances. OWL Full permet un maximum d'expressivité avec la liberté de syntaxe d'RDF. Il n'impose pas de séparation entre classe, propriété, individu et valeur des données. Il permet donc d'augmenter le sens du vocabulaire pré-défini (en OWL ou RDF). Il lève les contraintes imposées par OWL DL pour rendre certaines valeurs disponibles et utilisables dans des bases de données ou de connaissances, mais il ne supporte pas les raisonnements liés à la logique de description.

L'utilisation du langage de représentation OWL dans le cadre d'un processus de RI permet d'une part de faire reposer le SRI sur un standard mais surtout d'utiliser un langage incrémental. Dans un premier temps, les ontologies que nous considérons pourront être représentées à partir de OWL-Lite, puis elles évolueront vers un autre sous-langage lorsque le système sera capable de prendre en compte le niveau de formalisation spécifié.

4.4 Bilan

Les ontologies ont pour but de représenter un phénomène du monde à partir d'une « spécification explicite et formelle d'une conceptualisation ». Cependant, plusieurs types de ressources couvrant différentes structures et différents engagements sémantiques ont été définis pour rendre la représentation de la connaissance interprétable par l'homme et manipulable par la machine. Le degré de formalisation des ontologies dépend de l'engagement sémantique choisi, en d'autres termes, de la restriction faite sur l'interprétation de chaque concept (un concept renvoyant à un objet matériel, une notion ou une idée intervenant dans la spécification). Plus l'engagement sémantique choisi dans la formalisation est élevé, plus les problèmes liés à l'ambiguïté diminuent. Deux types de ressources peuvent être distingués à partir de la notion d'ontologie : les ressources terminologiques et les ressources conceptuelles. L'inconvénient des ressources terminologiques est qu'elles ne prennent en compte que le niveau lexical de la conceptualisation. Par leur manque de formalisation, elles laissent le champ libre à l'ambiguïté et à l'interprétation personnelle de leurs significations. Les ressources conceptuelles ont pour but de combler cette lacune en s'appuyant sur des formalisations plus ou moins définies. Elles reposent sur deux niveaux sémiotiques. Le niveau lexical couvre tous les termes ou labels définis pour transcrire le sens des concepts. Le niveau conceptuel représente les concepts et les relations conceptuelles entre eux. Les ontologies formelles sont manipulables et interprétables aussi bien par l'homme que la machine. Cependant, ce niveau de formalisation implique des contraintes de conception et de mise à jour. Bien que le degré d'engagement sémantique soit plus restreint dans le cas des hiérarchies de concepts et des ontologies dites légères, elles sont plus facilement conçues.

Les ontologies que nous retenons comme adaptées à la RI sont les ontologies légères de domaine à forte composante lexicale. Ces ontologies permettent d'intégrer l'ensemble des termes utiles pour la détection des concepts dans le corpus et présentent un degré de formalisation plus élevé que les ressources habituellement utilisées en RI (vocabulaire contrôlé, thésaurus, hiérarchie

de concepts). Elles offrent, de plus, des perceptives de traitements de l'information évolués pour le domaine. Les ontologies que nous considérons sont les ontologies de domaine car elles permettent de normaliser le sens des éléments considérés.

Parallèlement, de nombreux langages de représentation des ontologies ont été développés. Le langage le plus récent et polyvalent est le langage OWL car il s'inscrit dans le Web Sémantique et propose différents niveaux de formalisation permettant à la ressource représentée d'évoluer en fonction de l'engagement sémantique choisi. Nous choisissons donc ce langage. Il permet de faire évoluer la ressource en fonction du degré de formalisation qui pourra être géré par le système de RI. Ainsi, si l'intérêt d'utiliser des ontologies lourdes apparaissait, il pourrait facilement être intégré.

5 Conclusion

L'utilisation de connaissances dans les systèmes informatiques a pour but d'améliorer les mécanismes existants en accédant à la sémantique associée à l'information traitée.

L'utilisation de connaissances en RI vise à faciliter la modélisation du contexte de la recherche. Tout d'abord, les connaissances peuvent être utiles pour la compréhension du contenu des granules d'information en apportant une couche sémantique aidant à interpréter les mots qui le composent. Elles peuvent également aider à la prise en compte de la tâche de l'utilisateur, en reposant en particulier sur l'exploitation des méta-données qui sont associées aux granules, en explicitant leur rôle dans la description de ceux-ci. Enfin, elles peuvent également être utiles dans la compréhension du besoin de l'utilisateur aussi bien par l'utilisateur lui-même que par le système. Elles permettent de donner une vue générale sur la connaissance disponible dans un corpus et peuvent aider à spécifier le besoin d'information d'un utilisateur.

Dans ce contexte, notre choix s'est porté sur la représentation des connaissances à travers des ontologies. Plus spécifiquement, les ontologies que nous considérons sont des ontologies de domaine, car elles permettent de représenter la connaissance normalisée pour un domaine d'intérêt. Un des intérêts supplémentaires de notre choix réside dans le fait que les ontologies correspondent à des ressources conceptuelles formalisant un degré de connaissance plus élevé qu'une ressource terminologique. Elles doivent cependant intégrer une large composante lexicale afin que les concepts soient décelables dans les documents. Plus spécifiquement, nous utilisons deux types d'ontologies légères : une représente le domaine du thème de la collection étudiée, l'autre représente le domaine de la tâche que réalise l'utilisateur. L'accès à la collection par la navigation dans ces ontologies permet à l'utilisateur de se faire une idée sur la connaissance à sa disposition et le guide pour accéder à l'information qu'il recherche.

Avant de décrire l'utilisation des ontologies en RI, nous présentons dans le chapitre suivant les méthodes permettant de les concevoir.

Chapitre 2
Conception d'ontologies

1 Introduction

L'utilisation d'ontologies en informatique vise à intégrer une couche de connaissances aux systèmes afin de permettre des traitements élaborés de l'information qu'ils manipulent.

La conception d'ontologies est une tâche difficile qui nécessite 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 génération 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 dans le sens où nous les avons définies dans le chapitre 1 (ces ontologies se distinguent par la présence ou non d'axiomes). Cependant, ce procédé de génération est très coûteux en temps et pose surtout des problèmes de maintenance et de mise à jour [Ding 2002]. La conception automatique d'ontologies commence à émerger comme un sous-domaine de l'ingénierie des connaissances. Face à la masse croissante de documents présents sur le Web et aux avancées technologiques dans le domaine de la recherche d'information, de l'apprentissage automatique et du traitement automatique des langues, de nouveaux travaux portent sur la recherche d'un procédé plus automatique de génération d'ontologies. Ce mécanisme mène généralement à la conception d'ontologies dites légères. Dans [Maedche 2001], 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.

La RI manipule des documents textuels et des thésaurus ; notre étude se porte donc sur la conception d'ontologies à partir de ces deux types de ressources. Nous concentrons donc cet état de l'art sur les méthodes reposant sur ces éléments-là.

La deuxième section décrit les méthodologies et méthodes relatives à la construction d'ontologies à partir de textes et de thésaurus. La troisième section présente les techniques de mises à jour d'ontologies existantes. La dernière section est consacrée à la description des travaux visant à faire évoluer un thésaurus vers une ontologie.

2 Construction d'ontologies à partir de textes

La construction d'ontologies à partir de textes repose sur des méthodologies et des méthodes permettant de les mettre en œuvre. Elles sont décrites dans les sections suivantes.

2.1 Méthodologies de conception d'ontologies

2.1.1 Conception manuelle d'ontologies

Plusieurs principes ont été définis pour la construction d'ontologies [Gruber 1993] [Uschold 1996] [Guarino1998a]. Ces principes insistent sur la nécessaire clarté de la définition des éléments que l'ontologie doit contenir (rôle et portée de l'ontologie, définition des concepts, limitation des ambiguïtés) ainsi que sur la séparation des phases de conception et d'implantation dans un langage formel de l'ontologie. L'ensemble de ces principes reste cependant abstrait.

Des méthodologies ont également été définies pour cadrer le développement d'ontologies de domaine.

Dans le projet TOVE [Gruninger 1995a], l'ontologie de domaine est construite 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 [Fernandez 1997] et KACTUS (modelling Knowledge About Complex Technical systems for multiple USe) [Schreiber, 1995] 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 est intéressant dans la mesure où il évite de construire une ontologie à partir de rien. Cette problématique existe d'ailleurs dans divers domaines comme dans la conception des systèmes d'information avec la définition de patrons [Rieu 1999]. 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). Les différentes étapes proposées sont les suivantes :

- La première étape, étape de spécification, permet de produire un document de spécification de la future ontologie. Ce document décrit, entre autres, l'objet de l'ontologie, ses utilisateurs, ses utilisations, le degré de formalisation à employer,
- La deuxième étape d'acquisition de connaissances mène à l'identification des termes de l'ontologie et leur définition. Des techniques d'acquisition de connaissances, comme les réunions de brainstorming, les interviews d'experts, les analyses de textes, sont listées. Toute technique d'acquisition est donc a priori utilisable,
- L'étape suivante de conceptualisation vise à structurer la connaissance du domaine en un modèle conceptuel. La représentation est à ce stade informelle,
- L'étape d'intégration qui suit, permet d'envisager quelles sont les ontologies existantes qui pourraient être intégrées dans la construction de l'ontologie. Des ontologies génériques ou de haut niveau peuvent être utilisées comme structuration des concepts de base. D'autres ontologies peuvent également être utilisées pour la définition de termes communs,
- La phase suivante est celle de l'implantation. Elle consiste à représenter formellement l'ontologie à partir de langage comme LOOM, Ontolingua mais aussi Prolog ou C++,
- La phase d'évaluation intervient alors pour vérifier et valider l'ontologie en question ainsi que son environnement logiciel et sa documentation. Les problèmes de cohérence, d'incomplétude et de répétition sont alors vérifiés,
- La dernière étape repose sur la documentation. Les auteurs insistent sur ce point en précisant que, généralement, les documentations sont incomplètes pour la compréhension globale de l'ontologie et sa réutilisation. Ils proposent alors de pallier ce manque en imposant la rédaction de documentations à la fin de chacune des phases de la construction de l'ontologie.

La particularité de la méthodologie METHONTOLOGY est de s'attâcher fortement à la maintenance de l'ontologie de domaine et à son évaluation. De plus, les auteurs insistent sur le fait que les concepteurs doivent s'efforcer autant que possible d'utiliser des ontologies existantes. Nous nous appuyons sur cette méthodologie pour concevoir en particulier l'ontologie de domaine de la tâche (cf chapitre 4).

Un autre type de méthodologie vise à aider à la construction d'ontologie formelle en proposant des règles pour évaluer la cohérence logique des liens de subsomption modélisés [Guarino 2002] [Kassel 2002]. La méthodologie OntoClean [Guarino 2002] repose sur la définition de caractéristiques des concepts pour structurer des ontologies en imposant certaines contraintes sur l'utilisation des liens de subsomption [Guarino 2000]. Ces caractéristiques aussi appelées méta-propriétés sont :

- l'*identité* : un concept porte une propriété d'identité si cette propriété permet de conclure quant à l'identité de deux instances de ce concept. Cette propriété peut porter sur des attributs du concept ou sur d'autres concepts. Par exemple, le concept « étudiant » porte une propriété d'identité liée au « numéro » de l'étudiant, deux étudiants étant identiques s'ils ont le même numéro ;
- la *rigidité* : une propriété d'un concept est rigide si elle est essentielle pour chacune des instances du concept, c'est-à-dire que chacune des instances détient cette propriété pour exister. Par exemple, la propriété « être un humain » est rigide, mais « être un étudiant » est non rigide ;
- l'*unité* : un concept composé de plusieurs concepts est un concept unité si chacune de ses instances forme « un tout ». Par exemple, le concept « eau » n'est pas unité car une de ses instances est une quantité d'eau qui ne peut pas être reconnue en tant qu'entité isolée. Le concept « océan » est unité car ses instances telles que l'« océan atlantique » sont des entités à part entière.
- la *dépendance* : un concept C1 est dépendant d'un concept C2 si, pour toute instance de C1, il existe une instance de C2 qui ne soit ni partie ni constituant de l'instance de C1. Par exemple, « parent » est un concept dépendant de « enfant » (et inversement), car l'existence d'un parent suppose celle d'un enfant. Mais « couteau » et « manche » ne sont pas dépendants, car le manche fait partie du couteau.

Ces quatre méta-propriétés font peser des contraintes sur les liens de subsomption entre concepts. Par exemple, un concept portant une propriété d'identité ne peut subsumer un concept qui n'en porte pas. La méthodologie OntoClean propose donc de typer les concepts d'une ontologie à l'aide de ces caractéristiques, puis de tester la cohérence de la hiérarchie des concepts en vérifiant que les contraintes induites par ces caractéristiques ne sont pas violées. La méthodologie OntoSpec [Kassel 2002] reprend ces méta-propriétés et incite le concepteur à utiliser certaines propriétés en particulier lors de l'élaboration de l'ontologie. Elle prend en amont un ensemble d'entités conceptuelles exprimées par des termes ainsi qu'un ensemble de définitions en langue naturelle. Par un processus de transformation reposant sur la définition des différentes méta-propriétés, OntoSpec permet d'élaborer des ontologies semi-informelles pouvant ensuite être représentées dans le langage formel choisi par le concepteur.

Ces deux dernières méthodologies présentent l'avantage de fournir un cadre formel pour l'élaboration d'ontologies. Les ontologies que nous considérons n'ont pas besoin d'un tel niveau de formalisation. Notons cependant qu'il sera primordial de prendre en compte les méta-propriétés sur lesquelles elles reposent afin de faire évoluer les ontologies légères que nous considérons en ontologies formelles.

Différents outils ont été proposés pour aider à la conception manuelle d'ontologies. Ces outils permettent d'éditer une ontologie, d'ajouter des concepts et des relations, etc. Ils intègrent différents langages de formalisation (RDF, OWL). Certains doivent être installés en local alors que d'autres sont distribués sur le Web. Ces outils sont décrits plus spécifiquement.

OntoEdit (Ontology Editor) [Sure 2002] est un environnement de construction d'ontologies. Il a été développé par la compagnie Ontoprise. Il permet l'édition des hiérarchies de concepts et de relations dans le cadre de la logique des frames, ainsi que l'expression d'axiomes algébriques. Des outils graphiques dédiés à la visualisation d'ontologies sont inclus dans l'environnement. OntoEdit intègre, dans sa version commerciale, un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs ainsi qu'un plug-in permettant le test de la cohérence d'une ontologie. OntoEdit gère de nombreux formats de représentation de connaissances dont OWL, RDFS et FLogic.

Protégé2000 est une interface modulaire, développée au Stanford Medical Informatics de l'Université de Stanford, permettant l'édition, la visualisation, le contrôle (vérification des contraintes) d'ontologies [Noy 2000]. Le modèle de connaissances de Protégé2000 est issu du modèle des frames et contient des classes (concepts), des slots (propriétés) et des facets (valeurs des propriétés et contraintes), ainsi que des instances des classes et des propriétés. Protégé2000 autorise la définition de métaclasses, dont les instances sont des classes, ce qui permet de créer son propre modèle de connaissances avant de bâtir une ontologie. De nombreux plug-in sont disponibles ou peuvent être créés par l'utilisateur. Parmi ceux-ci, citons le plug-in permettant d'utiliser le langage OWL et les plug-ins de visualisation.

L'outil ODE, Ontology Design Environment, développé à l'Université de Polytechnique de Madrid permet de mettre en place la méthodologie METHONTOLOGY décrite dans la section 2. Son successeur pour le Web WebODE [Aspirez 2003] a pour ambition de couvrir l'ingénierie ontologique à travers les différentes activités liées au cycle de vie d'une ontologie : acquisition de connaissances à partir du Web, édition d'ontologies, test de la consistance d'une ontologie, alignement et fusion d'ontologies, import et export dans des formats variés. Le modèle de représentation de connaissances utilisé associe un modèle de type frame (concepts et attributs) avec des relations entre concepts. Des propriétés conceptuelles (en particulier algébriques) peuvent être associées aux relations. Les axiomes d'une ontologie sont des tautologies du domaine, mais on peut aussi inclure dans l'ontologie des règles susceptibles d'être utilisées pour raisonner dans un moteur d'inférence de type Prolog.

WebOnto, développé au Knowledge Media Institute de l'Open University, offre une interface graphique d'édition collaborative, de tests et de parcours d'ontologies sur le Web [Domingue 1998]. Le modèle de connaissance utilisé est celui du langage OCML (Operational Conceptual Modeling Language), un langage à base de Frames.

OilEd (Oil Editor) est un éditeur d'ontologies utilisant le formalisme DAML+OIL et les Logiques de Description [Bechhofer 2001]. Il est essentiellement dédié à la construction d'ontologies dont on peut ensuite tester la cohérence à l'aide de FACT, un moteur d'inférences bâti sur OIL. Il permet l'export d'ontologies sous les formats RDF, DAML+OIL, OWL et d'autres langages moins consensuels comme SHIQ.

ONTOLINGUA, développé au Knowledge Systems Laboratory de l'Université de Stanford, est un serveur d'édition d'ontologies permettant la construction collaborative d'ontologies [Farquhar 1997]. Une ontologie y est directement exprimée dans un formalisme également nommé ONTOLINGUA.

2.1.2 Construction d'ontologies de domaine à partir de textes

La méthodologie TERMINAE [Aussenac 2000a] [Aussenac 2000b] est une méthode qui repose sur l'analyse de corpus linguistique. Elle tente de répondre aux manques des autres méthodologies en proposant une approche pour sélectionner les concepts, leurs propriétés, les relations et leur regroupement. Elle repose pour cela 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 seront réalisés. Cette étape est fondamentale car de la qualité du corpus dépendra la qualité des traitements. Le corpus doit couvrir entièrement le domaine

traité par l'application. Cette phase nécessite l'intervention d'un expert pour récolter les différents types de documents significatifs,

- La troisième étape correspond à l'étude linguistique. Des outils sont utilisés sur le corpus afin d'extraire les termes et leurs relations lexicales et syntaxiques. Le choix des outils est laissé à l'utilisateur. Une application de la méthode est proposée par les auteurs à partir des outils LEXTER [Bourigault 1996] et Caméléon [Seguela 1999]. Le premier extrait les termes candidats à partir de leurs dépendances syntaxiques. Le second extrait des relations entre termes à partir de patrons linguistiques. Les outils utilisés nécessitent l'intervention d'experts du domaine afin de sélectionner et de valider les candidats. A la fin de cette étape, un ensemble de termes, de relations lexicales entre ces termes et de regroupements est obtenu,
- 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. Des méthodes de formalisation telle que celle définie dans [Kassel 2002] peuvent être utilisées dans cette étape.

Cette méthodologie 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. Cependant, elle ne spécifie pas comment les concepts doivent être sélectionnés ni quelles sont les propriétés adéquates. Cette méthodologie est générale et demande l'intervention d'experts. Nous nous appuyons sur cette méthodologie pour définir une méthode d'élaboration d'ontologies à partir de textes intégrant les ressources terminologiques disponibles telles que les thésaurus (cf chapitre 5).

Un outil est associé à cette méthodologie. Il a été développé au LIPN de l'Université Paris-Nord2 et permet, à travers les outils LEXTER [Bourigault 1996], Syntex [Bourigault 2000]et Caméléon [Seguela 1999], d'extraire d'un corpus textuel les candidats termes d'un domaine. Il offre un support méthodologique qui permet de faire évoluer progressivement une ontologie en conservant des liens entre les textes et les niveaux linguistiques et conceptuels. Plus de détails sont fournis dans [Bourigault 2003] et [Aussenac 2002]. Le modèle de représentation de TERMINAE est celui des Logiques de Description mais une traduction des ontologies dans le langage OWL est possible.

2.2 Méthodes de construction d'ontologies de domaine à partir de textes

Une méthodologie définit le cadre général de la conception d'ontologies mais nécessite l'ajout de méthodes pour la mettre en œuvre concrètement.

La construction d'ontologies à partir de textes vise à cette mise en œuvre à partir d'éléments qui peuvent être extraits de ces textes. Elle aboutit généralement à la conception d'ontologies légères de domaine. Dans les sections suivantes, nous décrivons les différentes méthodes issues de différents domaines (RI, IC, traitement automatique des langues) qui peuvent être associées aux phases principales des méthodologies.

La première phase décrit la constitution d'un corpus de référence. La deuxième phase concerne l'extraction des termes et des concepts du domaine. Les deux phases suivantes extraient les relations taxonomiques et associatives de la structure de l'ontologie. Enfin, des techniques permettent de mettre à jour une ontologie. Ces différentes phases sont décrites dans les sections suivantes.

2.3 Constitution du corpus

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 lequel reposera cette élaboration [Condamines 2005].

Plusieurs cas de figures peuvent amener à élaborer ce corpus [Lame 2002].

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 est d'interroger le Web à partir de requêtes décrivant le domaine qui devra être traité dans l'ontologie. Dans [Agirre 2000] par exemple, l'objectif est de mettre à jour WordNet. Pour cela, un ensemble de documents de référence est extrait du Web pour chacun des concepts à mettre à jour à partir de requêtes formées des termes qui décrivent le concept et ses hyperonymes dans WordNet. 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 [Ok 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.

2.4 Extraction de termes

Les termes candidats pour représenter les concepts d'une ontologie peuvent être extraits selon deux approches : syntaxique ou statistique. L'approche syntaxique analyse le rôle grammatical des mots dans ces textes, alors que l'approche statistique repose sur la fréquence d'apparition des mots dans les textes.

2.4.1 Techniques syntaxiques d'extraction de termes

Les techniques syntaxiques extraient des termes à partir des relations grammaticales entre les mots dans les phrases des documents. Ces termes peuvent être composés d'un seul mot ou d'une suite de mots. Les expressions extraites syntaxiquement, aussi appelées syntagmes, exploitent le rôle des mots dans les documents dont elles sont issues. Elles déterminent des composants de la phrase très précis qui doivent être détectés en fonction de la grammaire de la langue utilisée. Elles contiennent un ou plus d'un mot et sont plus petites qu'une phrase [Caropreso 2000]. On peut extraire des expressions nominales telles que «le toit de la maison », des expressions verbales : « jouent au patin à glace», « vont à l'école », des expressions adjectivales : « immédiatement disponible ». Les expressions extraites syntaxiquement prennent en compte des relations linguistiques entre les mots, elles sont donc plus significatives sémantiquement que la juxtaposition des mots formant les expressions statistiques.

Différents analyseurs syntaxiques existent. [Riloff 1996] s'appuie sur l'utilisation de patrons syntaxiques définis manuellement. Nomino [David 1990], quant à lui, repose sur le découpage des textes en unités lexicales pour l'identification de syntagmes nominaux. Syntex [Bourigault 2000] s'appuie sur un apprentissage des relations de dépendance entre mots pour extraire les syntagmes de différents types (nominaux, verbaux,…) et les organiser suivant un réseau de dépendance. La particularité de Syntex est de s'appuyer sur un apprentissage endogène

du corpus qui lui permet d'être plus performant qu'un analyseur reposant uniquement sur des règles définies manuellement. Il peut s'adapter à des collections spécialisées dans différents domaines tels que le droit et la chirurgie [Bourigault 2002a] [Le Moigno 2002]. Nous l'avons donc utilisé dans nos travaux dans le cadre d'une coopération avec l'Equipe de Recherche en Syntaxe et Sémantique de Toulouse qui a développé ce logiciel. Nous décrivons en détail son fonctionnement.

2.4.1.1 Syntex

Syntex repose sur un pré-traitement du corpus par un étiqueteur pour préciser la nature des mots du corpus (adjectif, verbe, …) puis sur une analyse syntaxique pour extraire des syntagmes.

Pré-traitement par un étiqueteur : identification du rôle des mots

Avant d'être analysé par Syntex, le corpus est traité par un étiqueteur, aussi appelé positionneur morpho-syntaxique. Ce dernier permet d'obtenir la nature de chacun des termes présents dans le corpus (adjectif, verbe, nom,…).

- Pour le français, le logiciel Cordial (réalisé par la société Synapse Développement[1]) est utilisé. Au-delà de la correction orthographique et grammaticale, ce logiciel offre de nombreuses fonctionnalités : analyse syntaxique visualisable, analyse sémantique, statistiques, recherche d'occurrences et de mots clés. La deuxième édition du logiciel ajoute une option supplémentaire : un étiqueteur morpho-syntaxique utilisé par Syntex.
- Pour l'anglais, Syntex utilise le TreeTagger. Il repose sur la construction d'un arbre de décision binaire pour estimer les probabilités d'obtenir un certain rôle syntaxique. Contrairement aux autres étiqueteurs reposant sur les modèles de Markov, il n'a pas besoin d'un large ensemble d'apprentissage [Schmid 1994].

Les étiqueteurs s'appuient sur des dictionnaires orthographiques et sur des dictionnaires grammaticaux pour réaliser leur analyse.

Analyse syntaxique : extraction des relations de dépendance

Dans un premier temps, pour chaque phrase, Syntex procède à l'identification des constituants syntaxiques maximaux (verbaux, nominaux, adjectivaux) que détermine la structuration en relation de dépendance. Pour chaque mot recteur, il construit un syntagme maximal en parcourant toutes les relations de dépendance syntaxique dont ce mot est la cible jusqu'à aboutir à des mots qui, soit ne sont pas recteurs, soit sont têtes d'un syntagme maximal déjà construit. Syntex identifie pour cela les principales relations de dépendance qui sont : sujet de verbe, objet de verbe, complément prépositionnel de verbe, de nom ou d'adjectif, épithète du nom. Il détermine également des relations plus complexes en retrouvant l'antécédent d'un pronom relatif ou le rattâchement d'une préposition.

L'ensemble des relations de dépendance se décrit sous la forme d'un triplet (recteur, relation, régi) ; les relations étant orientées du régi vers le recteur. On peut citer, par exemple, le triplet (français, sujet, perd) représentant la relation « sujet » dans la phrase « le français perd de l'importance » ; le recteur est ici « français » et le régi « perd ».

[1] http://www.synapse-fr.com/

Un module différent est écrit pour extraire du texte chacune des relations de dépendance. Chaque module repose sur un traitement en deux étapes.

- Recherche des candidats de la relation de dépendance

  Selon le module considéré, l'analyse consiste à trouver le recteur ou le régi de la relation. Dans le cas du rattâchement d'une préposition par exemple, le module recherche le recteur, c'est-à-dire les verbes, les noms et les adjectifs qui apparaissent dans le contexte gauche de la préposition et qui sont susceptibles de la régir. De manière à établir la liste des candidats, des règles reposant sur la syntaxe sont écrites à la main pour décrire les différentes configurations possibles.

- Sélection du candidat de type *régi*

  Lorsqu'un seul candidat a été détecté par la première étape, il est considéré comme valide et est mémorisé par l'analyseur.

  Dans le cas où une ambiguïté est présente entre plusieurs candidats, l'analyseur calcule un indice de plausibilité pour chacun d'eux. Ceci est réalisé grâce à un apprentissage endogène. Ce principe est la particularité de Syntex. Il permet à l'analyseur de se servir du corpus et des relations qu'il a déjà établies. Ainsi, il bénéficie de plus d'informations syntaxiques et lexicales pour identifier des relations complexes ou ambiguës. Il utilise la nature des candidats et le rôle qu'ils ont dans les relations précédemment décelées. En fonction d'une stratégie de choix, un des candidats est retenu. Syntex a plus de facilité à s'adapter à des contextes spécialisés qu'un analyseur reposant uniquement sur des règles prédéfinies.

  Les syntagmes ainsi extraits sont dits « syntagmes maximaux ».

Analyse syntaxique : extraction d'un réseau de dépendance

A partir des relations des syntagmes maximaux extraits à l'étape précédente, Syntex extrait l'ensemble des syntagmes et les organise suivant un réseau de dépendance. Pour cela, il se base sur la structure d'un syntagme. La caractérisation de la structure d'un syntagme est la suivante :

- une tête, qui est constituée du mot recteur avec sa catégorie ;
- une liste de couples (relation, expansion), chaque expansion étant (le lemme d') un mot régi ou (la forme normalisée d') un syntagme dont la tête est un mot régi, et la relation étant la relation de dépendance syntaxique ;
- une forme normalisée, constituée à partir du lemme de la tête et de la séquence des lemmes ou formes normalisée des expansions.

Le réseau de dépendance est construit en ajoutant pour chaque syntagme maximal différent rencontré : (1) un nœud dont le label est la forme normalisée du syntagme, (2) des liens vers les nœuds correspondant à ses expansions, étiquetés par le nom de la relation de dépendance. Puis il complète la liste des syntagmes en ajoutant les réductions et les simplifications des syntagmes maximaux. Les réductions sont obtenues en diminuant le nombre d'expansions que peut contenir le syntagme jusqu'à ce qu'il ne contienne plus qu'une relation unique liant deux expressions. Par exemple, le syntagme verbal maximal « français perd faculté de répudiation » sera réduit en deux syntagmes : « français perd » et « perd faculté de répudiation ». Les simplifications consistent à remplacer, dans une relation, une expression par sa tête. En reprenant l'exemple précédent, « français perd faculté de répudiation » sera simplifiée par « français perd faculté ». La simplification s'applique également sur les expressions réduites. Ainsi « perd faculté » sera également ajouté à la liste.

2.4.1.2 Sélection des syntagmes

Les syntagmes extraits par un analyseur syntaxique qui sont les plus représentatifs du corpus, et donc du domaine, peuvent être sélectionnés par rapport à leur nature grammaticale (verbe, syntagme verbal, syntagme nominal,…) ou par rapport à leur poids dans le corpus. Ce poids peut être calculé par des mesures issues de la lexicologie comme, par exemple, la mesure d'information mutuelle [Velardi 2001]. Cette mesure (IM) est utilisée pour les syntagmes composés de plusieurs termes et sélectionne les syntagmes dont les termes sont les plus liés. Elle est définie par :

$$IM(x,y)=\frac{nb(x,y)}{nb(x)*nb(y)}$$

Avec nb(i), le nombre d'apparitions du terme i dans la collection, et nb(i,j) le nombre d'apparitions du terme i avec le terme j dans un même contexte (le contexte d'apparition est généralement le document) et x, y deux termes composant un même syntagme.

Alternativement, des mesures statistiques peuvent être utilisées. Ces mesures sont décrites dans la section 2.4.2.2. Les candidats extraits syntaxiquement sont ensuite validés par un expert [Le Moigno 2002].

2.4.2 Techniques statistiques d'extraction de termes

2.4.2.1 Extraction des termes

L'extraction de termes se base sur l'utilisation d'un anti-dictionnaire pour supprimer les mots vides puis sur la radicalisation des termes restants pour supprimer les variantes lexicales [Risjbergen 1979].

L'utilisation d'un anti-dictionnaire vise à éliminer les mots ayant un contenu informationnel vide. Ces mots apparaissent dans la plupart des documents et ne sont pas discriminants. Ces mots qui peuvent être des articles, prépositions, conjonctions voire même des verbes sont appelés mots vides et sont regroupés dans un anti-dictionnaire. L'utilisation d'un anti-dictionnaire permet de réduire considérablement le nombre de termes extraits. Par exemple, si les termes de l'anti-dictionnaire apparaissent dans plus de 80 % des documents, le nombre de termes diminue de 40% [Baeza-Yates 1999]. Afin de supprimer les différentes variantes lexicales d'un terme et de ne considérer qu'une forme unique, la racine du terme est extraite. Ce procédé est appelé radicalisation [Frakes 1992]. Il existe différentes méthodes de radicalisation : l'utilisation de tables de correspondance entre le terme et le radical, l'utilisation de n-grammes[2] et la suppression de suffixes ou de préfixes [Lovins 1968] [Porter 1980]. Ainsi, l'extraction de termes individuels (composés d'un seul mot) est la plus utilisée. Cependant, l'extraction d'expressions permet d'obtenir des termes ayant une meilleure sémantique. Les expressions extraites statistiquement représentent une séquence de mots juxtaposés. Seules les expressions présentes fréquemment dans la collection sont extraites [Mitra 1997], [Sebastiani 1999]. Une fois sélectionnées, elles peuvent être également radicalisées. Chacun des termes de l'expression est mis sous sa forme radicalisée et les termes de l'expression sont, par exemple, ordonnés alphabétiquement [Mitra 1997]. Ces techniques, bien que définies dans le cadre de l'indexation de documents pour la RI, peuvent être appliquées dans le cadre de la construction d'ontologies.

2.4.2.2 Sélection des termes

Les termes sont ensuite sélectionnés à partir de leurs occurrences dans les documents. Des mesures issues de la RI peuvent être utilisées. Dans ce cas, la fréquence du terme dans la

[2] un n-gramme est un suite de n caractères

collection permet d'extraire les termes très utilisés. La mesure tf.idf (voir chapitre 3 pour plus de détails) [Robertson 1976] permet de prendre en compte les termes qui apparaissent souvent dans le corpus mais principalement dans quelques documents. L'entropie, quant à elle, analyse la répartition des termes dans les documents, ce qui permet, en fonction de la formule, de sélectionner les termes soit rares soit redondants [Brini 2005].

Plusieurs conclusions contradictoires ont été tirées sur l'efficacité comparative de ces différentes mesures. Les résultats présentés dans [Ok Koo 2003] pour l'extraction de termes du domaine de l'économie à partir de documents du Wall Street Journal montrent que la fréquence des termes donne de meilleurs résultats que la mesure tf.idf. Les auteurs de [Maedche 2004] préconisent l'utilisation de la mesure tf.idf. Les travaux présentés dans [Lame 2002] sur l'application de ces mesures pour la sélection de syntagmes du droit n'ont pas permis de déterminer la « meilleure approche », ou bien comment les combiner. L'auteur explique ces résultats par la spécificité de son corpus de référence et la difficulté de déterminer les termes du droit, domaine qui couvre le monde et auquel tous les objets peuvent appartenir.

Dans [Velardi 2001] une mesure a été définie pour extraire des termes propres à un domaine. La mesure définie, appelée Pertinence au Domaine (PD), prend en compte la probabilité d'obtenir un terme à la fois dans le corpus spécialisé et dans des corpus généraux et d'autres domaines.

Soit un ensemble (D1, D2, …, Dn) composé de n domaines. La pertinence pour le domaine D_i du terme t est calculée par :

$$PD(t,Di)=\frac{P(t/Di)}{\sum_{i=1..n} P(t/Di)}$$

où une estimation de $P(t/D_i)$ est $E(P(t/Di))=\frac{freq(t \text{ in } Di)}{\sum_{i=1..n} freq(t \text{ in } Di)}$

Cette mesure favorise les termes n'apparaissant que dans le domaine considéré.

Les termes d'un domaine D_i sont ensuite sélectionnés pour représenter un consensus dans le domaine. Ce facteur est défini à partir de la mesure CD analysant la distribution du terme t sur un ensemble de documents d_j appartenant au domaine D_i. Plus un terme est utilisé dans le domaine, plus il est jugé représentatif de ce domaine.

$$CD(t,Di)=\sum P(t,dj)\log_2\left(\frac{1}{P(t,dj)}\right)$$

où une estimation de $P(t/d_j)$ est $E(P(t/dj))=\frac{freq(t \text{ in } dj)}{\sum_{i=1..n} freq(t \text{ in } dj)}$

Ces mesures ont été testées afin d'extraire les termes du domaine économique et les termes du tourisme [Velardi 2001]. Les termes extraits ont été comparés aux termes contenus dans des dictionnaires ou thésaurus du domaine. Les résultats montrent que tous les termes représentatifs du domaine ne sont pas extraits (F-mesure[3] de 30% dans la première

[3] F-mesure = $\frac{2*Rappel*Précision}{Rappel+Précision}$

expérimentation et un rappel[4] de 2% dans la deuxième) mais que ceux qui l'ont été sont majoritairement pertinents (précision[5] de l'ordre de 80% dans les deux expérimentations).

Aucune étude n'a conclu sur le choix entre une extraction syntaxique ou statistique des expressions en fonction de l'application (indexation de documents, construction d'ontologies). Ce choix reste donc difficile. Cependant, intuitivement, on peut penser que pour la construction d'ontologies l'extraction syntaxique est la plus adaptée. Nous retenons de plus que les mesures de sélection des termes dépendent fortement du domaine traité et du corpus de référence confectionné. L'intervention d'experts pour la validation des termes est nécessaire.

2.5 Extraction de liens de subsomption

Les liens de subsomptions dans une ontologie permettent d'organiser les concepts hiérarchiquement.

Pour extraire ce type de liens, différentes méthodes issues de la RI et de l'IC existent. En réalité, certaines d'entre elles permettent d'extraire des liens non pas entre concepts mais entre termes et les liens obtenus ne respectent pas la définition stricte de la subsomption (cf chapitre 1 section 4.2). Elles s'appuient sur des approches statistiques ou linguistiques. Les approches statistiques regroupent et structurent les termes par rapport à leurs occurrences dans les différents documents. Les approches linguistiques reposent sur une analyse syntaxique du contenu des documents et regroupent les termes ou des concepts par rapport à leur contexte d'apparition.

2.5.1 Approches statistiques

Nous décrivons plusieurs méthodes statistiques permettant d'extraire des relations taxonomiques entre termes. Ces méthodes se basent sur l'analyse des co-occurrences entre termes dans les documents. La co-occurrence correspond à l'apparition simultanée de deux termes dans un texte (document ou fenêtre de N mots). Les méthodes présentées dans cette section représentent l'ensemble des co-occurrences dans une matrice. Cette matrice est ensuite utilisée pour regrouper hiérarchiquement les termes, soit par application des méthodes de classification automatique, soit en s'appuyant sur des mesures de probabilité.

2.5.1.1 Méthodes de regroupement hiérarchique de termes

[Manning 1999] indique que le regroupement non supervisé en classes permet la détection de relations de généralisation. Ainsi, [Maedche 2000] propose d'appliquer les méthodes hiérarchiques ascendantes ou descendantes à la matrice de co-occurrence des termes extraits des documents. Dans le cas d'une classification hiérarchique ascendante, au départ, chaque classe est composée d'un terme. Les regroupements se font en associant deux classes qui ont les profils les plus proches. Le profil d'une classe correspond à ses occurrences avec l'ensemble des termes. Les regroupements successifs aboutissent à la formation d'une seule classe. La classification hiérarchique descendante procède, au contraire, par divisions successives. Elle part d'une classe formée de tous les termes et la divise en deux classes. Elle détermine la classe la moins cohérente, au sens de la mesure de similarité entre classes et la divise elle aussi en deux. Ceci est répété jusqu'à l'obtention des partitions d'un seul élément. Dans les deux cas, de nombreuses mesures de similarités ont été définies ou utilisées dans la littérature [Jain 1999] [Kullback leiber 1951] [Murtagh 1998].

---

[4] Rappel= $\frac{\textit{nombre de termes corrects extraits}}{\textit{nombre de termes corrects}}$

[5] Précision = $\frac{\textit{nombre de termes corrects extraits}}{\textit{nombre de termes extraits}}$

2.5.1.2 Méthode reposant sur la probabilité de co-occurence

Dans [Sanderson 1999], l'association des termes repose sur une relation parent-enfant où le terme parent est plus général que le terme enfant. Cette relation entre termes est extraite d'après la co-occurrence asymétrique de termes. La relation est caractérisée par les deux règles suivantes :

$$p(x/y) \geq O.8(\textit{seuil empirique}) \text{ et}$$

$$P(y/x) < P(x/y)$$

où p(x/y) est la probabilité d'obtenir le terme x dans un document sachant que le document contient le terme y, inversement pour p(y/x).

La première règle assure que les deux termes apparaissent suffisamment dans les mêmes documents (en l'occurrence dans 80% des cas). D'après la deuxième règle, x susbsume y si les documents dans lesquels il apparaît sont un sous-ensemble des documents où apparaît x. Le terme apparaissant le plus souvent est choisi comme parent. Les relations extraites à partir des deux règles citées sont ensuite nettoyées en supprimant les termes qui co-occurrent dans moins de deux documents et en supprimant les relations redondantes par rapport à la propriété transitive de la relation. Si les relations a subsume b, a subsume c et b subsume c sont extraites, la relation a subsume c peut être supprimée parce qu'elle est déductible des deux autres.

2.5.2 Approches linguistiques

Les approches linguistiques s'appuient sur une analyse syntaxique des documents pour extraire des relations taxonomiques entre termes issus des documents. L'analyse syntaxique permet soit de définir des patrons d'extraction, soit de procéder au regroupement conceptuel.

2.5.2.1 Approches reposant sur la définition de patrons d'extraction

L'idée d'utiliser des patrons lexico-syntaxiques sous la forme d'expressions régulières afin d'extraire des relations sémantiques a été introduite par Hearst [Hearst 1992]. Dans ces travaux, les patrons syntaxiques définissant la relation taxonomique sont construits manuellement. Les relations sont ensuite extraites automatiquement du corpus. Un exemple de patron est le suivant « *SN {,SN}*{,}ou autres SN* » où SN dénote la présence d'un syntagme nominal. En appliquant ce patron sur l'extrait suivant « préfectures, mairies ou autres bâtiments publics », les relations taxonomiques sont déduites entre les couples bâtiments public/préfecture, bâtiment public/mairie.

Le Système Prométhée développé par Morin propose d'étendre ces travaux en se basant sur un apprentissage permettant d'extraire automatiquement les patrons lexico-syntaxiques correspondant à une relation sémantique donnée [Morin 1999]. L'apprentissage consiste tout d'abord à donner au système un ensemble de couples de termes vérifiant cette relation. Un corpus est ensuite analysé et l'ensemble des patrons que suivent ces couples est traité par le système. Les patrons sont triés et sélectionnés à partir d'une mesure de similarité permettant de choisir les patrons les plus représentatifs de l'ensemble. Les patrons retenus sont ensuite validés par un expert. Le système permet également d'étendre les relations déduites entre termes simples (un seul mot) à de nouvelles relations entre termes composés de plusieurs mots dont deux des mots sont liés. Si, par exemple, les termes *pomme* et *fruit* sont liés sémantiquement, le même lien sémantique pourra être déduit entre les termes composés « *jus de pomme* » et « j*us de fruit* ». Cette étape, appelée normalisation sémantique, repose sur l'analyse syntaxique des syntagmes et restreint les cas où la formation des termes composés est possible. Une relation sémantique entre les mots t_1 et t_2 n'implique pas toujours une relation entre les syntagmes t_1t_1, et t_2t_2, où les t_i sont des mots. Une relation sémantique est déduite entre les syntagmes t_1t_1, et t_2t_2, si les trois règles suivantes sont respectées :

- 1 : Une relation sémantique lie (t_1 et t_2) ou ($t_{1'}$ et $t_{2'}$) et les mots non sémantiquement liés sont identiques ou sont morphologiquement liés.

- 2 : t_1 et $t_{1'}$ sont deux mots têtes et t_2 et $t_{2'}$ sont deux arguments avec des rôles thématiques semblables.

- 3 : t_1 $t_{1'}$ et $t_2t_{2'}$ partagent la même relation sémantique entre leurs composants.

Les auteurs se basent sur une hypothèse qui dit que si 1 et 2 sont vrais alors 3 l'est aussi. Afin de mettre en place les règles 1 et 2, une analyse morphologique des termes est réalisée à partir de l'outil d'acquisition FASTER [Jacquemin 1999]. Les rôles thématiques sont extraits à partir de la nature des termes et des prépositions les encadrant. Un ensemble de patrons, réalisés manuellement à partir des résultats du logiciel d'acquisition ACABIT [Daille1996], définit les rôles thématiques semblables. Voici un exemple d'application. Une relation peut être déduite entre *composé chimique de la graine* et *composition du fruit* car tout d'abord les termes *fruit* et *graine* sont donnés comme étant liés et les termes *composé* et *composition* sont morphologiquement similaires. De plus, un patron syntaxique permet de déterminer que les deux syntagmes sont des arguments du même rôle thématique. Le système a été testé pour la détection de relations taxonomiques à partir d'un ensemble de couples d'apprentissage extraits de documents ou bien d'un thésaurus [Morin 1999a]. Les résultats par les deux apprentissages sont similaires et montrent qu'environ 60% des relations taxonomiques détectées à partir des patrons appris sont correctes et qu'à peu près 80% des relations étendues par la normalisation sont justes. Le système pourrait permettre d'extraire d'autres types de relations.

Les patrons ont aussi été utilisés dans [Maedche 2000] pour créer une ontologie à partir de dictionnaires dans le domaine de l'assurance et des télécommunications. Les auteurs présentent leur méthode comme étant plus originale que la précédente car les patrons sont générés au niveau des concepts et non pas au niveau des termes. Pour cela, ils considèrent les concepts comme étant le premier mot de la définition des entrées d'un dictionnaire. Les patrons sont ensuite définis à partir de ce terme.

2.5.2.2 Regroupements conceptuels

Faure et Nedellec [Faure 1998] ont présenté le système ASIUM. Dans ce système, les relations taxonomiques sont acquises par un traitement syntaxique des documents. Des classes sont formées à partir des termes qui apparaissent après le même verbe et la même préposition en appliquant un algorithme de regroupement conceptuel. Les classes sont successivement agrégées pour former de nouveaux concepts et une hiérarchie constituant l'ontologie. Les classes formées doivent être labellisées par un expert pour identifier le concept qu'elles représentent. Les classes sont composées de groupements de mots suivant le patron: <verbe> <rôle syntaxique |préposition : nom*>*, comme par exemple « <voyager> <sujet : humain> <par : véhicule>. Les couples <rôle syntaxique : nom> ou <préposition : nom> sont appelés « mots têtes ». La mesure de similarité qui permet d'évaluer la distance entre les classes, et donc de les regrouper, dépend de la proportion de mots têtes communs dans les différentes classes en prenant en compte leur fréquence d'apparition dans les documents. La méthode a été testée sur un corpus de recettes de cuisine. Lorsque le système est entraîné à retrouver les couples verbe-argument sur 30 % du corpus, la hiérarchie proposée est valide à 30%.

[Assadi 1999] décrit un cas pratique de la détection de relations à partir de regroupements de syntagmes reposant sur des critères lexicaux et conceptuels appliqués au domaine de l'électricité. Les syntagmes sont regroupés à partir des modifieurs (prépositions, adjectifs, verbes) qui les suivent dans le contexte des documents.

Dans OntoLearn [Velardi 2001], des sous-graphes conceptuels sont créés pour faciliter la génération manuelle de taxonomies. Les sous-graphes sont formés en regroupant les syntagmes ayant la même tête et en développant la hiérarchie au fur et à mesure que des termes sont ajoutés à la queue des syntagmes.

2.6 Détection de relations non taxonomiques

Une autre phase dans l'élaboration d'ontologies consiste à extraire des relations non taxonomiques entre concepts. La difficulté est qu'elle doit non seulement extraire des relations entre concepts mais également permettre de labelliser les relations afin de désigner la relation sémantique. Les approches présentées ici sont des contributions qui rentrent dans ce cadre. Il s'agit d'approches statistiques reposant sur la co-occurrence et d'approches syntaxiques.

2.6.1 Co-occurrence des verbes

[Ok Koo 2003] propose de construire une ontologie pour aider à la Recherche d'Information. Cette ontologie porte sur le domaine de l'économie ; elle est construite semi-automatiquement à partir de l'analyse des documents du Wall Street Journal des collections TREC. Le principe consiste à extraire les noms et noms propres apparaissant fréquemment. Ces termes sont appelés les termes centraux. Afin de déceler des relations entre termes, les termes apparaissant dans une fenêtre de quatre mots autour des termes centraux sont extraits. Les termes co-occurrant fréquemment sont proposés pour être reliés dans l'ontologie, les verbes apparaissant dans le contexte sont proposés pour être labels de la relation.

2.6.2 Analyse syntaxique

Plusieurs approches reposant sur l'analyse syntaxique des documents permettent d'extraire des relations non taxonomiques.

L'approche présentée dans [Velardi 2001] consiste à analyser syntaxiquement les documents et à produire des triplets (Acteur Procédé Objet) à partir du patron syntaxique (Sujet Verbe Objet). Les couples vérifiant ce patron sont sélectionnés si au moins un élément du triplet appartient déjà à l'ontologie. Les couples ayant une faible plausibilité sont supprimés. La plausibilité prend en considération le pouvoir consensuel des termes sur le domaine décrit dans la section 3.2. Aucun résultat n'est donné sur les performances de la méthode.

[Bourigault 2002b] propose également des mécanismes pour extraire des relations (non typées). Le principe utilisé s'appuie sur la proximité entre les différents syntagmes extraits par Syntex et sur l'analyse distributionnelle [Harris 1968]. Cette analyse consiste à regrouper des syntagmes en fonction du contexte (mots par lesquels ils sont régis et mots qu'ils régissent) qu'ils partagent. Dans [Bourigault 2002], les syntagmes déduits de l'analyse syntaxique sont rapprochés s'ils sont formés autour de la même relation et de la même tête. La proximité est calculée ensuite grâce à différentes formules qui prennent en compte la "productivité" d'un contexte (c'est-à-dire le nombre de termes qui apparaissent dans ce contexte), et la "productivité" d'un terme (nombre de contextes différents dans lesquels apparait le terme). L'analyse distributionnelle permet donc de rapprocher des termes deux à deux en fonction des contextes qu'ils partagent mais aussi de rapprocher les contextes en fonction des termes qu'ils ont en commun. Le module UPERY reposant sur cette analyse a été intégré à Syntex. Ce module permet donc de rapprocher des termes par rapport à leur contexte et de détecter une relation sémantique entre eux. Cependant, la nature de cette relation n'est pas spécifiée.

2.6.3 Approche reposant sur les règles d'association

[Maedche 2000] et [Sugiua 2004] proposent d'extraire des relations non taxonomiques entre concepts à partir de règles d'association. L'algorithme d'exploration de données présenté dans [Srikant 1995] permettant de trouver des associations généralisées entre éléments tels que les

supermarchés et les achats des clients est utilisé. Le principe de l'algorithme consiste à déterminer un ensemble de transactions $T := \{t_i / i=1..n\}$ où chaque transaction t_i est constituée d'un ensemble d'éléments $X_i := \{a_{ij} / j=1..m,\ a_{ij} \in C\}$ et chaque élément a_{ij} référence un concept appartenant à l'ensemble C. Dans le cas de la conception d'ontologies, les éléments sont les termes. Dans [Maedche 2000], les transactions retenues sont l'apparition de groupes de noms ou d'entités nommées dans un même syntagme (contenant une préposition ou bien apposant les mots). L'approche proposée a été évaluée sur un corpus des télécommunications et permet d'extraire des relations intéressantes du point de vue des auteurs. En revanche, dans [Sugiua 2003], les transactions sont le fait de retrouver deux mêmes termes dans une phrase. Testée sur un corpus du droit, cette méthode permet d'obtenir des relations suivant des taux de rappel et de précision relativement bas (respectivement 27% et 24%). L'algorithme [Srikant 1995] génère des règles d'association $X_k \Rightarrow Y_k$ telles que le support et la confiance de ces règles excèdent un seuil fixé. Le support de l'association $X_k \Rightarrow Y_k$ est défini par le pourcentage de transactions qui contient $X_k \cup Y_k$. La confiance correspond au pourcentage de transactions où Y_k apparaît dans la transaction si on a X_k. Cet algorithme est initialement étendu pour prendre en compte les ancêtres dans une taxonomie des éléments $a_{i,j}$. Le support et la confiance sont alors calculés pour les transactions où Y_k ne contient aucun ancêtre de X_k. Puis les règles d'association obtenues sont filtrées pour ne contenir aucune règle d'association faisant intervenir $X_k \Rightarrow Y_k$ $\underline{X_k} \Rightarrow \underline{Y_k}$ où $\underline{X_k}$ est un ancêtre de X_k et $\underline{Y_k}$ est un ancêtre de Y_k. Cependant cette méthode ne permet pas de fournir des labels aux relations extraites.

Les travaux présentés dans [Maedche 2000] ont été étendus dans [Kavalec 2004] afin d'extraire les labels. Les labels sont extraits parmi les verbes co-occurrant fréquemment autour des deux concepts c_1 et c_2. Des expérimentations ont été menées sur le corpus SemCor. Ce corpus présente l'avantage que chacun des termes est désambiguïsé à partir d'un sens précis qui peut être retrouvé dans WordNet. L'utilisation de ce corpus a permis de regrouper les termes co-occurrents à partir des concepts les généralisant dans WordNet. Les résultats obtenus montrent que la moitié des labels proposés permettent de désigner correctement les relations sémantiques entre concepts.

2.7 Bilan

Les approches proposées dans la littérature aident à extraire des relations taxonomiques et associatives ; elles se basent sur des analyses syntaxiques ou statistiques du corpus. Ces méthodes se placent pour la plupart au niveau des termes et non pas au niveau des concepts. L'extraction du niveau conceptuel reste une problématique de la construction d'ontologies. L'extraction de relations taxonomiques est étudiée depuis plusieurs années et offre différentes solutions. En revanche, l'extraction de relations associatives apparaît comme une préoccupation récente. Le principal problème qui demeure est la proposition de labels pour ces relations.

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

l'utilisateur. Onto-Builder autorise aussi la fusion d'ontologies extraites de différents sites Web. Pour une étude comparative détaillée des outils de conception d'ontologies, nous renvoyons le lecteur aux travaux suivants [Su 2002] [Sure 2003].

Comme le montrent les sections précédentes, de nombreuses méthodes et outils ont été proposés dans le domaine de la RI et dans le domaine de l'IC permettant d'extraire des termes ou de détecter des relations entre eux. Il faut toutefois noter que la majorité de ces méthodes élabore la connaissance en partant uniquement des textes. Pourtant, des ressources terminologiques ou conceptuelles peuvent exister ; il est alors dommage de ne pas les considérer, même si elles ne sont pas complètes. Nous voyons dans la section suivante comment une ontologie peut être mise à jour en s'appuyant sur un corpus de référence.

3 Techniques de mise à jour d'ontologies

Différentes techniques de mise à jour de l'ontologie ont été proposées. Ces techniques visent à extraire de nouveaux termes et à les intégrer dans l'ontologie. Elles se basent sur la détection d'indices lexicaux et statistiques liant les nouveaux termes détectés dans le corpus et les labels de concepts présents dans le corpus. Plusieurs approches ont été définies pour extraire ces indices.

La méthode proposée dans [Faatz 2002] consiste à définir une mesure calculant la distance dans les textes entre les labels de concepts de l'ontologie liés par une propriété et d'utiliser cette mesure pour détecter de nouveaux labels. Les labels extraits des documents sont alors supposés être liés par la propriété 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 sémantique considérée. La mesure est optimisée à partir des labels des concepts liés par cette relation dans l'ontologie. Deux corpus de référence sont utilisés pour extraire les nouveaux termes. Un premier corpus est constitué sur le Web par la spécification de requêtes à partir des labels des concepts dans l'ontologie. Le deuxième corpus est composé de documents du domaine considéré. Le cas d'application présenté dans l'article consiste à définir une mesure permettant d'extraire la relation sémantique « est un ». Cette propriété 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 (seuls les termes co-occurrant au moins deux fois sont considérés). La méthode est testée dans le domaine de la médecine à partir d'une ontologie spécialisée. Sept concepts de l'ontologie sont considérés pour optimiser la mesure. Les résultats obtenus sur le corpus constitué de documents issus du Web permettent d'obtenir un nombre plus restreint de nouveaux labels que ceux provenant du le corpus spécialisé. Cependant, l'analyse de ces différents labels montre que ceux proposés à partir du premier corpus sont plus pertinents. Bien qu'elle n'ait été testée que pour la relation « est un », 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 tous types de relations est difficile car elle implique de maîtriser les indices du corpus qui permettent de déceler la relation.

Dans [Maedche 2002a], une autre méthode est définie. Cette méthode vise à détecter de nouveaux termes mais propose uniquement de les intégrer à l'ontologie à partir de relations taxonomiques. Une matrice de co-occurrence est constituée à partir de termes désignant les concepts de l'ontologie et les termes co-occurrant autour d'eux dans une fenêtre de trois mots dans une même phrase. Ces termes sont ensuite hiérarchisés à partir d'une méthode hiérarchique ascendante et d'une méthode hiérarchique descendante décrite dans la section 3.3.1.1. 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 par les méthodes hiérarchiques sont proposés pour être ajoutés à l'ontologie comme sous-concepts des concepts à partir des labels avec lesquels ils sont liés. L'évaluation a consisté à prendre une ontologie existante portant sur le domaine du tourisme et à supprimer des concepts puis d'essayer de les retrouver et de les replacer correctement. Le nombre moyen de réponses correctes ainsi que le degré d'erreur de placements sont comparés aux résultats obtenus par l'algorithme des k plus proches voisins sans prendre en compte l'ontologie existante. L'algorithme de classification des k plus proches voisins a l'avantage de ne mettre en place aucun apprentissage et considère qu'un nouvel objet doit être ajouté à la classe existante dont il est le plus similaire parmi les k plus proches. Les résultats ne permettent pas d'établir d'amélioration des performances en prenant en compte l'ontologie existante. Les concepts pères sont seulement mieux retrouvés à partir de l'algorithme ascendant. Les résultats obtenus par cette méthode ne permettent pas de mettre en valeur son intérêt.

Un autre type d'approche vise à mettre à jour les ontologies à partir de la méthode des signatures de thématique [Lin & Hovy 2000], [Hovy & Lin 1999]. Habituellement utilisée pour la génération 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 réaliser la signature de concepts. La méthode proposée dans [Agirre 2000] vise à mettre à jour WordNet par rapport à un corpus donné, en supprimant les sens inutiles des concepts et en en proposant de nouveaux, 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 contenus dans les synsets ainsi que les hyperonymes trouvés dans WordNet. 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 la signature du concept. Les termes de la signature sont retenus comme labels du concept. Les termes qui sont partagés par différentes signatures permettent de trouver des liens entre les concepts. Dans [Agirre 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 à partir de leur signature mais ne permet pas de déterminer comment ces liens doivent être interprétés et où les concepts doivent être ajoutés dans l'ontologie.

Dans [Alfonseca 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-occurrant autour des termes définissant les concepts dans WordNet. La méthode permet donc d'affiner l'ontologie en déterminant où il faut ajouter les nouveaux concepts. 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 termes issus de WordNet sont retenus pour être 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 génération d'ontologies car elles permettent de spécifier le sens des concepts.

4 D'un thésaurus vers une ontologie

Faire migrer les thésaurus vers le Web Sémantique présente plusieurs avantages. Les thésaurus sont des ressources largement utilisées dans le domaine de la Recherche Documentaire. Ces ressources ont demandé de lourds efforts de conception et représentent un consensus sur les termes utilisés dans le domaine qu'elles représentent. Elles sont, de plus, largement utilisées par les documentalistes. Les thésaurus ont été jusqu'ici représentés à partir de normes dédiées et stockés sous format texte, format XML ou dans des bases de données, éventuellement accessibles à partir d'interfaces HTML. Transcrits dans des formalismes tels que RDF et OWL, ils seront uniformément représentés et pourront être distribués sur le Web. Les outils de visualisation et d'exploration pourront être utilisés sur ces nouvelles versions des thésaurus. Leur utilisation pour l'annotation de documents dans le cadre du Web Sémantique (voir chapitre 3) est, de plus, un enjeu du domaine. Plusieurs initiatives ont été menées pour proposer un modèle de représentation des thésaurus dans les langages dédiés au Web Sémantique. Un exemple de schéma et un exemple de méthode sont proposés dans la section 4.1.

Une alternative consiste à faire migrer les thésaurus vers le Web Sémantique non plus en les représentant dans un langage dédié mais en les transformant en ontologies. Cependant, les thésaurus ont été créés pour aider les documentalistes dans leur tâche d'indexation et de formulation de requêtes. L'utilisation à des fins d'indexation automatique est limitée par leurs caractéristiques décrites dans [Soergel 2004]. Les thésaurus manquent d'un niveau d'abstraction conceptuelle. Ce sont des collections de termes qui sont organisés suivant une seule hiérarchie ou bien plusieurs (cf les thésaurus à facettes) avec des relations basiques entre termes. La distinction entre un concept et sa lexicalisation n'est pas clairement établie. Les thésaurus ne reflètent pas comment le monde peut être compris en terme de sens. La couverture sémantique des thésaurus est, de plus, limitée. Les concepts ne sont généralement pas différenciés par rapport à leur type abstrait (comme par exemple les substances, les procédés). Les relations entre termes restent vagues et ambiguës. La relation « est lié à » est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques. Il est souvent difficile de déterminer les propriétés des relations « plus spécifique », « plus générique » qui peuvent regrouper les relations « est une instance de » ou « est une partie de ». Les thésaurus manquent également de consistance et peuvent contenir des informations contradictoires. Les thésaurus sont finalement difficilement exploitables dans des procédés automatiques à cause de ces différentes limites. Les ontologies légères ou lourdes permettent de pallier ces manques.

Une des difficultés majeures dans la migration des thésaurus vers des ontologies est de capturer la sémantique implicitement présente dans ces ressources utilisées habituellement par des personnes.

Dans la section 4.2 nous présentons différents projets portant sur la transformation de thésaurus en ontologies.

4.1 Migrer les thésaurus vers le Web Sémantique

Dans le but de faire cohabiter les ontologies formelles et les thésaurus, plusieurs méthodes ont été proposées pour représenter les thésaurus à partir des langages dédiés au Web Sémantique.

La figure 2.1 rappelle les relations entre termes les plus utilisés dans les thésaurus. Pour plus de détails, voir la section 3.2.2.4 du chapitre 1.

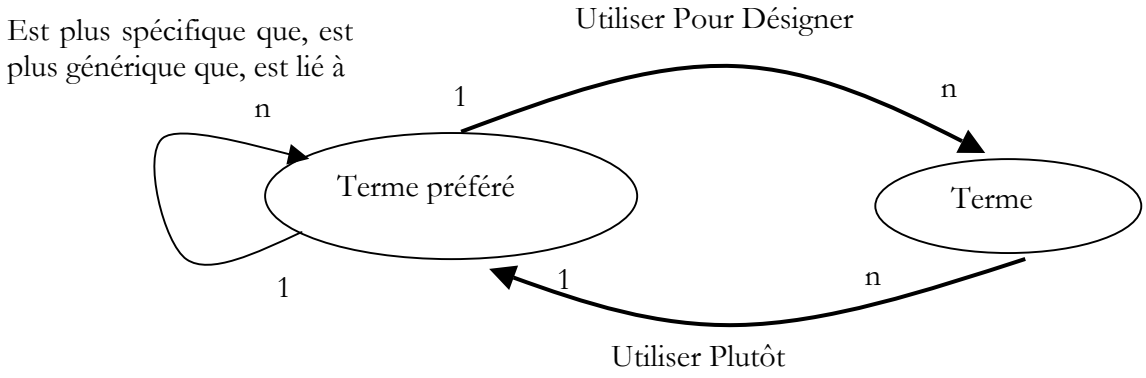


Figure 2.1 : Rappel des relations entre termes dans un thésaurus

Le projet mené par la SWAD-E (W3C Semantic Web Advanced Development for Europe) sur la migration des thésaurus propose un schéma pour publier les thésaurus sur le Web Sémantique [Miles 2003]. Ce projet repose sur la transformation du thésaurus en transcrivant littéralement la structure et les données initialement présentes. Le schéma ne prévoit pas d'ajouts d'informations pour capturer l'information implicitement présente dans les thésaurus. Le schéma SKOS proposé est orienté concept pour des facilités de manipulation et de maintenance. Il repose sur les propriétés et concepts de RDF et RDF-S.

- Chaque « terme préféré » devient un label préféré d'un concept,
- Les « termes non préférés » deviennent des labels alternatifs du concept,
- Un label peut être une chaîne de caractères, un symbole ou une image,
- Les relations de termes « est plus spécifique », « est plus générique » ou « est lié à » deviennent des relations sémantiques entre les concepts,
- Les concepts peuvent avoir des annotations telles que des notes les décrivant ou des définitions.

Un exemple de transformation d'un extrait du thésaurus standard (figure 2.2) est présenté dans la figure 2.3. Cet exemple est tiré de [Miles 2004].

Therapy
NT Back care

Back care
BT Therapy
RT Back pain

Back pain
UF Backache
RT Back care
Musculoskeletal disorders

Musculoskeletal disorders
UF Bone disorders
RT Back pain

Figure 2.2 : Extrait d'un thésaurus standard

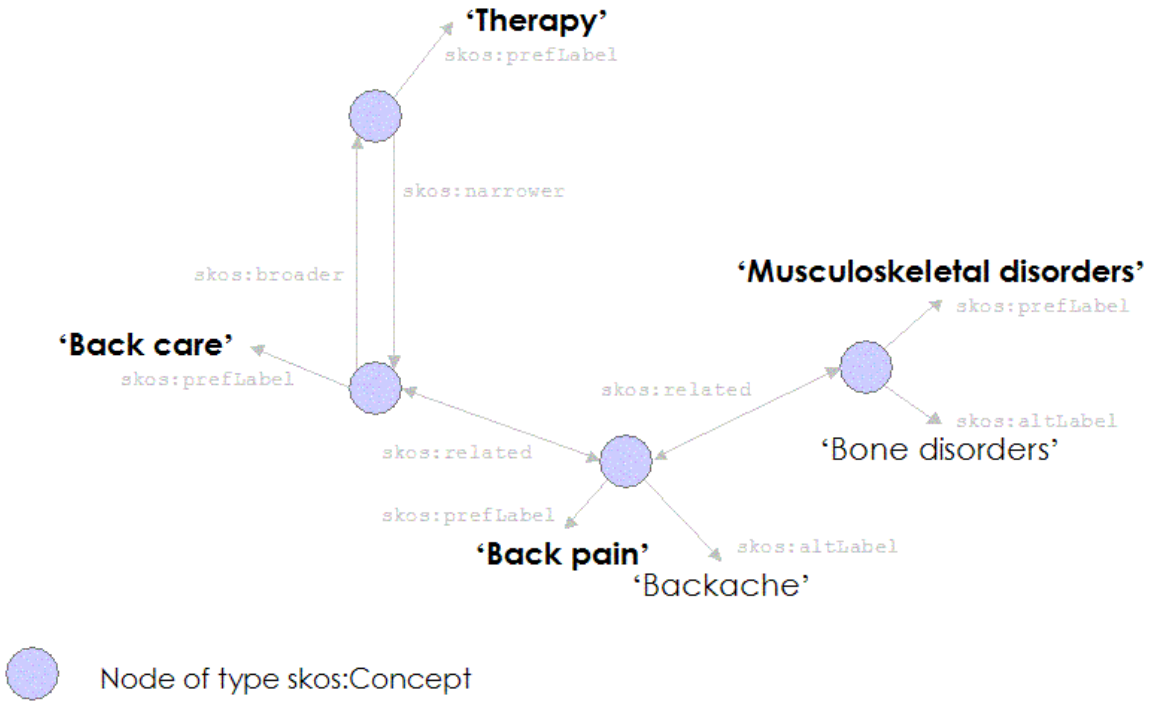


Figure 2.3 : Transformation du thésaurus à partir de Skos Schéma

Récemment, une nouvelle version de SKOS schéma vient d'être proposée [Miller 2005]. Cette version intègre des éléments de OWL tels que les propriétés qui peuvent être définies sur les relations (transitivité, symétrie…). La propriété de transitivité est ajoutée aux relations « plus spécifique », « plus générique », de même que la symétrie à la relation « est lié à ». Nous notons cependant que dans un contexte général de transformation de thésaurus, ces ajouts peuvent poser problème car la nature de ces relations n'est pas uniforme. Par exemple, dans certains thésaurus, la relation « est plus spécifique » peut englober la relation « est une instance », pour laquelle la relation de transitivité ne peut pas s'appliquer (voir chapitre 1). Plusieurs propriétés ont été ajoutées à la classe « label ». La propriété « label caché » permet d'ajouter à un concept des labels décrivant les variations lexicales que peuvent prendre les termes le décrivant. Cette propriété est dite cachée car elle peut être utilisable par la machine pour détecter le concept mais cachée à

l'utilisateur qui connaît ces variations. Une autre sous-classe de labels permet de spécifier les labels symboliques correspondant à des images décrivant les concepts.

Le schéma proposé présente l'avantage de dissocier la représentation d'un concept et de ses labels. Ses labels peuvent de plus être de différentes natures (terme principal, variation lexicale ou image …). Cependant, un schéma seul ne suffit pas ; une méthodologie permet de définir les étapes de la transformation du thésaurus dans le schéma choisi.

Une méthodologie a été proposée dans [Van Assem 2004] pour assurer la migration de thésaurus en RDF/OWL. Elle a pour but de s'adapter à n'importe quelle spécification de thésaurus. Les thésaurus comportant d'autres relations que les relations les plus utilisées présentées dans la figure 1 sont aussi considérés. Cette méthode se veut donc générale et propose de capturer l'information implicitement présente dans le thésaurus. Elle repose sur trois étapes principales : la transformation syntaxique, la transformation sémantique et la standardisation. La transformation syntaxique se décompose en deux étapes. La première consiste à conserver la structure initiale du thésaurus en le transformant à partir d'éléments simples de RDF (définition des classes : rdf :class, définition des noms des classes rdfs:labels) et des types de données supportés par XML (tels que les entiers : xsd:integer, les dates, …). Cette étape doit se faire en préservant le sens des informations dans le thésaurus (rajouter si nécessaire des commentaires : rdfs:comment, des références vers des documentations : rdfs:seeAlso), en évitant la redondance d'information et l'interprétation sur les données du thésaurus. La deuxième étape vise à expliciter la structure syntaxique de la future ontologie en ajoutant de nouveaux éléments qui étaient sous-entendus dans la représentation originale et qui ont besoin d'être formalisés dans la représentation conceptuelle. Par exemple, la mise en évidence du terme principal pour représenter un ensemble de synonymes peut se faire en mettant en gras le terme dans la représentation texte ; dans l'ontologie, cette distinction devra être intégrée en créant par exemple une nouvelle classe de termes. La transformation sémantique se décompose elle aussi en deux étapes. La première est l'explicitation de la sémantique contenue dans la version initiale comme, par exemple, l'ajout de caractéristiques aux propriétés (transitivité : owl:TransitiveProperty, symétrie : Owl:SymmetricProperty). La deuxième est l'introduction d'interprétation comme, par exemple, l'ajout d'une nouvelle classe permettant de regrouper plusieurs classes comme étant ses filles. La dernière étape de standardisation consiste à lier le schéma de modélisation de l'ontologie proposée à un schéma standard visant à représenter un méta modèle de thésaurus. Cette étape reste délicate car aucun modèle ne présente jusqu'ici un consensus [Van Assem 2004].

Cette méthodologie permet la séparation des différentes phases menant à la transformation du thésaurus. Ceci présente l'intérêt de limiter l'interprétation des éléments représentés dans le thésaurus et mène à la construction d'un nouveau thésaurus respectant la connaissance initialement représentée. Cependant aucun outil n'est développé pour aider à sa mise en œuvre.

D'autres méthodes ont été proposées pour permettre la transformation d'un thésaurus en une ontologie. Elles sont décrites dans la section suivante.

4.2 Raffinement de thésaurus en ontologies

Différents projets portent sur la transformation de thésaurus en ontologies. Ces projets ne visent pas uniquement à transformer les thésaurus dans un langage du Web Sémantique mais à formaliser la connaissance qu'ils contiennent.

Dans [Wielinga 2001], une approche est présentée afin de transformer le thésaurus à facettes de l'art et de l'architecture AAT en ontologie pour indexer des images. Cette approche est entièrement manuelle. L'ontologie est formalisée en RDFS. La première étape a consisté à identifier les concepts. Les concepts sont représentés par deux propriétés. La première indique le label utilisé pour représenter le terme principal dans le thésaurus initial. La deuxième propriété

regroupe l'ensemble des synonymes du terme dans le thésaurus. Les concepts sont organisés hiérarchiquement à partir des 33 hiérarchies qui organisent les termes principaux dans le thésaurus à facettes. La deuxième étape augmente les concepts à partir de nouvelles propriétés. Les propriétés sont définies pour des classes intermédiaires et identifient les propriétés qu'ont les classes filles. Par exemple, les concepts représentant des styles et des périodes sont augmentés par des propriétés telles que « période de temps à partir de », « période de temps jusqu'à ». De nouvelles classes sont également créées à partir d'une interface intégrée à la plate-forme Protégé 2000. Cette interface permet de définir des classes intermédiaires à partir de nombreuses propriétés et de les lier aux classes jusque-là présentes dans l'ontologie. Cette approche nécessite des traitements manuels et ne repose sur aucun apprentissage.

Dans [Soergel 2004], certaines étapes sont automatisées. Cette méthode repose sur la définition de «règles au fur et à mesure » (rules as you go). Elle a été appliquée à la transformation du thésaurus AGROVOC couvrant le domaine de l'agriculture, de la forêt, de la nourriture et des domaines reliés tel que l'environnement. Elle repose sur trois étapes. La première consiste à définir la structure de l'ontologie représentant le modèle conceptuel à un haut niveau d'abstraction. Le modèle conceptuel utilisé prend en compte les classes de concepts, termes et chaînes de caractère. La figure 2.4 définit les relations entre ces différents éléments. Chaque terme, concept, chaîne de caractères peut être identifié à partir d'une URI, ou d'un identifiant d'UMLS.

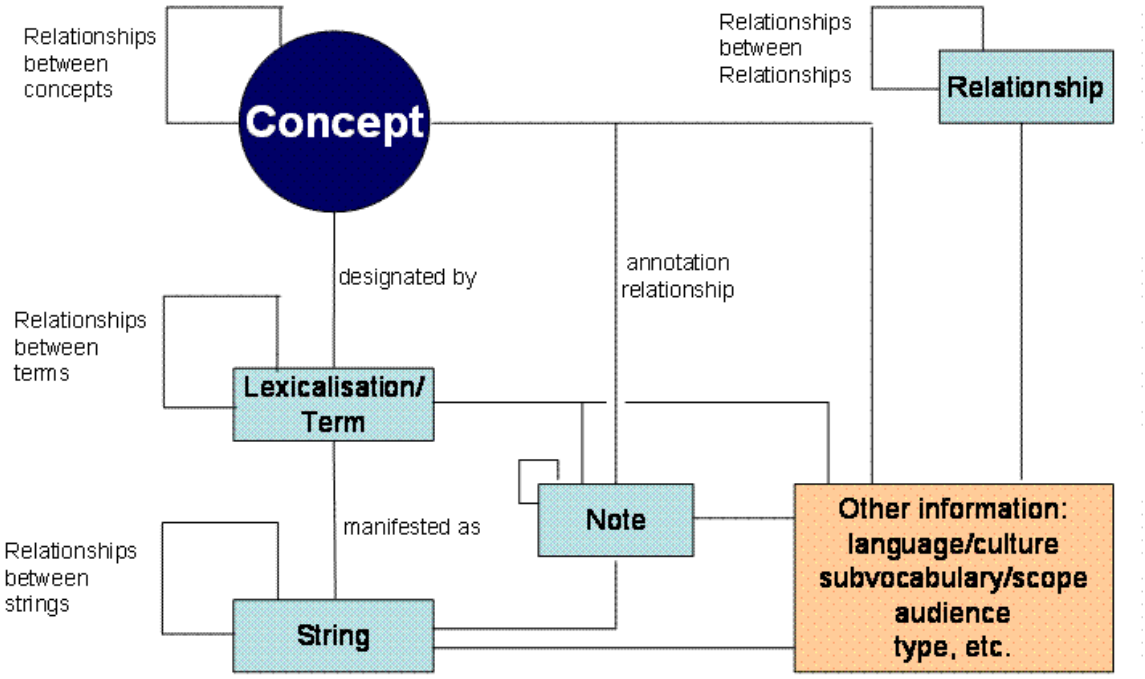


Figure 2.4 : Schéma conceptuel utilisé dans [Soergel 2004]

La deuxième étape consiste à identifier et lister les relations entre concepts pertinents pour le domaine représenté dans l'ontologie. Lors de cette étape, un extrait de couples reliés par chacune des relations du thésaurus est analysé afin de déterminer quelles sont les relations sémantiques qui peuvent en être déduites. Par exemple, dans le cas du thésaurus AGROVOC, les

relations « a comme ingrédient », « contient la substance » ont été déduites de la relation « est lié à ».

La troisième étape est l'extraction de ces relations. Cette étape est la plus longue. Elle se base sur une phase d'apprentissage qui demande à un expert d'évaluer un certain nombre de relations existant dans le thésaurus. A partir de patrons définis sur cet ensemble d'apprentissage, le système propose ensuite de nouvelles relations.

Un exemple de ce procédé est présenté dans la figure 2.5.

Un expert détermine que la relation

« vache NT lait de vache » doit se transformer en « vache A_COMME_COMPOSANT lait de vache ».

L'expert généralise ensuite ce patron à

« animal A_COMME_COMPOSANT lait ».

Le système propose alors que la relation

« chèvre NT lait de chèvre » doit se transformer en « chèvre A_COMME_COMPOSANT lait de chèvre ».

Figure 2.5 : Exemple de règle de généralisation

La phase d'apprentissage est très lourde pour l'expert qui doit, non seulement spécifier la relation sémantique induite entre les deux termes, mais également généraliser la relation et typer les concepts. Afin de limiter les patrons, la phase 2 limite les relations qui peuvent être déduites entre les termes par l'expert. Dans le projet, il a été montré que cette étape était très difficile à mettre en place ; de nombreuses erreurs impliquent la validation des résultats par un expert. De plus, comme certains patrons étaient impossibles à définir pour chacune des relations, une relation sémantique vague telle que « est lié à » a été maintenue dans l'ontologie finale pour regrouper les relations que le système est incapable d'analyser.

Les précédentes méthodes visent à transformer un thésaurus en une ontologie. D'autres méthodes vérifient la consistance de la connaissance représentée. [Hahn 2004] propose de créer une ontologie formelle de la médecine à partir du réseau sémantique UMLS. Le but de l'approche est de capturer la sémantique informelle de ce réseau créé pour l'indexation manuelle de ressources médicales et de réduire les définitions circulaires et inconsistantes qu'il contient. Les auteurs se focalisent sur un sous-domaine de la médecine qui est l'anatomie et les pathologies. L'originalité de leur approche est qu'ils se basent pour la formalisation sur la construction de triplets définissant les concepts faisant intervenir la relation « est un » mais aussi une sous-relation de « partie de » qui est « partie anatomique de ». La transitivité de ces deux relations est utilisée pour classifier les concepts en analysant la consistance de l'ontologie à partir d'un moteur d'inférence. Les concepts relatifs aux pathologies et à l'anatomie sont extraits de l'UMLS ainsi que les relations « est un », «est une partie de» et « est situé dans ». Les concepts extraits sont représentés sous forme de triplets de nœuds : un nœud représentant la structure de concept, un nœud représentant l'entité et un nœud représentant toutes les entités qui peuvent être rattachées au concept comme étant une de ses parties. Les concepts définis comme étant des parties du concept sont classés en dessous de ce dernier nœud. La relation « est située dans » est utilisée pour décrire les concepts. A partir de l'analyse des relations entre les différents triplets, les concepts sont classés et les cycles ou inconsistances identifiés. Ce procédé est uniquement utilisable sur des thésaurus qui contiennent des informations suffisamment élaborées mais il n'est pas applicable dans le cas de thésaurus standards dans lesquels seule la relation associative « est lié à » est présente.

4.3 Bilan

Plusieurs schémas et méthodes ont été proposés pour représenter un thésaurus dans un langage dédié au Web Sémantique. La caractéristique principale que nous retenons est la possibilité de faire évoluer la représentation obtenue vers une représentation plus formelle. OWL est le langage adapté à ce genre d'évolution par la spécification de ses trois sous-langages.

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

Nous proposons de transformer un thésaurus en s'appuyant sur la connaissance qu'il représente et sur les informations contenues dans un corpus documentaire du domaine.

5 Conclusion

La capture de la connaissance nécessaire pour l'élaboration d'une ontologie peut être réalisée à partir de plusieurs principes et méthodologies. Les différentes méthodologies proposées insistent sur l'importance de spécifier la tâche pour laquelle l'ontologie est construite. Cette tâche conditionne les éléments de connaissance qui devront être collectés ainsi que le niveau de formalisation nécessaire pour que l'ontologie soit manipulée par le système. Les méthodologies reposent généralement sur deux étapes : la spécification d'une ontologie informelle puis la formalisation de cette ontologie. Dans le cas où l'ontologie à construire est une ontologie lourde, cette seconde étape est longue car elle nécessite la spécification des différents axiomes décrivant la connaissance.

L'élaboration d'ontologies à partir de textes permet de faciliter la conception d'ontologies légères. Elle peut reposer soit sur une analyse statistique des termes apparaissant dans les documents, soit sur une analyse syntaxique qui consiste à analyser le rôle grammatical des mots qui les composent. Ces deux approches permettent tout d'abord d'aider à extraire les termes qui définiront le lexique de l'ontologie et qui seront les labels des concepts et des relations. Elles permettent également d'aider à définir la structure de l'ontologie à partir de l'extraction de relations taxonomiques et de relations non taxonomiques. Ces méthodes ne se basent sur aucune connaissance préalable du domaine. Pourtant, dans de nombreux domaines, des thésaurus ont été construits et ont nécessité de lourds efforts de conception.

Ainsi, la migration de thésaurus vers les technologies du Web Sémantique est un enjeu d'actualité. Largement utilisés dans le contexte de la recherche documentaire, les thésaurus sous formats dédiés au Web Sémantique pourront être distribués et exploités par les outils développés pour les ontologies. Ils pourront également être utilisés dans l'annotation des ressources du Web qui est une tâche importante du Web Sémantique. Cet aspect est décrit en détail dans le chapitre 3. La migration des thésaurus vers le Web Sémantique permet également d'envisager l'évolution de ces ressources terminologiques vers des ontologies. Cette évolution implique de capturer la connaissance implicitement présente dans les thésaurus habituellement utilisés par des documentalistes. Les approches de la littérature sont soit manuelles soit demandent un effort important à l'expert sollicité pour valider les connaissances élaborées pour l'apprentissage. Une

autre de leurs limites est de ne pas mettre à jour la connaissance qui a évolué depuis la création du thésaurus choisi.

Nous proposons de créer une ontologie à partir d'un thésaurus existant et en exploitant la connaissance issue d'un corpus de référence du domaine. L'information du corpus est utilisée non seulement pour spécifier des relations ambiguës entre termes mais également pour enrichir et mettre à jour l'ontologie résultante. Le processus que nous mettons en place est semi-automatique et vise à limiter au maximum les interventions de l'expert. L'ontologie construite est utilisée à des fins de RI.

Afin de spécifier les tâches auxquelles doivent répondre les ontologies dédiées à la RI, nous décrivons dans le chapitre suivant l'utilisation des ontologies en RI.

Chapitre 3
Utilisation des ontologies en RI

1 Introduction

Un des enjeux actuels de la RI est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'objectif est double : « comprendre » les contenus des documents et « comprendre » le besoin de l'utilisateur pour pouvoir les mettre en relation.

Les ontologies sont utilisées pour représenter des descriptions partagées et plus ou moins formelles de domaine et ainsi ajouter une couche sémantique aux systèmes informatiques. C'est donc naturellement que des travaux sur l'intégration des ontologies dans les SRI se développent. Une première solution vise à construire une ontologie à partir du ou des corpus sur lesquels les tâches de RI vont être réalisées [Saias 2003] [Ok Koo 2003]. Comme nous l'avons vu dans le chapitre 2, de nombreuses techniques permettent la conception d'ontologies. Cette solution assure a priori l'adéquation entre l'ontologie construite, le corpus et la tâche à réaliser. Cette solution reste cependant coûteuse et ne prend pas en compte l'existence de ressources qui pourraient être réutilisées. De plus, grâce à l'intérêt croissant pour les ontologies dans le domaine des systèmes d'information, de plus en plus d'ontologies sont maintenant accessibles. Une seconde solution est alors la réutilisation de ressources existantes. Dans ce cas là, les ontologies sont généralement choisies uniquement à partir du domaine de connaissance qu'elles abordent [Vallet 2005] [Baziz 2005] [Hearst 1997]. Nous considérons que ce choix doit être fait de manière plus approfondie. Les ontologies considérées doivent être adaptées aux tâches de RI visées et surtout apporter de la connaissance pertinente par rapport à l'information présente dans les corpus. Nous détaillerons dans une première section dans quelle mesure les ontologies peuvent être réutilisées ainsi que les travaux permettant d'analyser leur réutilisabilité et leur adéquation à un corpus.

Une fois l'ontologie choisie, la connaissance qu'elle représente peut être utilisée à différents niveaux dans le processus de RI. Elle peut aider à l'indexation des documents, alors appelée indexation sémantique. Ce point est présenté dans la section 3. Les ontologies peuvent également aider à la formulation du besoin de l'utilisateur et à l'accès aux documents. Ces aspects sont présentés dans la section 4. Enfin l'ontologie peut être utilisée dans le modèle lui-même pour réaliser l'appariement entre le besoin et les granules documentaires. Cette thématique sera développée dans la section 5. Plusieurs de ces processus intègrent des mesures de similarités entre concepts de l'ontologie. Nous voyons cet aspect dans la section suivante.

2 Similarités entre concepts dans une ontologie

L'évaluation du lien sémantique entre deux concepts dans une ontologie est un problème de longue date dans le domaine de l'intelligence artificielle et de la psychologie. La similarité sémantique est une évaluation du lien sémantique entre deux concepts dont le but est d'estimer le degré par lequel les concepts sont proches dans leur sens [Resnik 1999]. La définition donnée par Lin de la similarité sémantique repose sur trois suppositions [Lin 1998]. La similarité entre deux concepts est liée aux caractéristiques qu'ils ont en commun (plus ils ont de caractéristiques communes, plus les concepts sont similaires) et à leurs différences (plus deux concepts sont différents, moins ils sont similaires). La similarité maximale est obtenue lorsque deux concepts sont identiques.

La majorité des travaux portant sur le calcul de similarité dans une ontologie considèrent que la similarité peut être évaluée uniquement à partir des liens taxonomiques (ou lien « est un ») [Rada 1989] [Resnik 1995] [Wu 1994] [Jiang 1997]. D'autres, au contraire, estiment que ce calcul doit intégrer les autres types de liens [Lord 2003] [Thieu 2004]. Nous développons ces deux approches ainsi que leur évaluation dans les sections suivantes.

2.1 Similarité dans une taxonomie

Différentes mesures ont été définies pour permettre le calcul de la similarité entre concepts. Ces mesures sont classées par rapport aux caractéristiques des concepts permettant d'évaluer la similarité. Ces caractéristiques reposent soit sur la distance entre les concepts à travers leurs liens dans l'ontologie, soit sur l'information contenue par les concepts, soit sur les deux.

L'exemple de hiérarchie de concepts présenté dans la figure 3.1 est utilisé pour illustrer les différentes mesures. Les concepts sont représentés par des rectangles et les flèches symbolisent la relation « est un ».

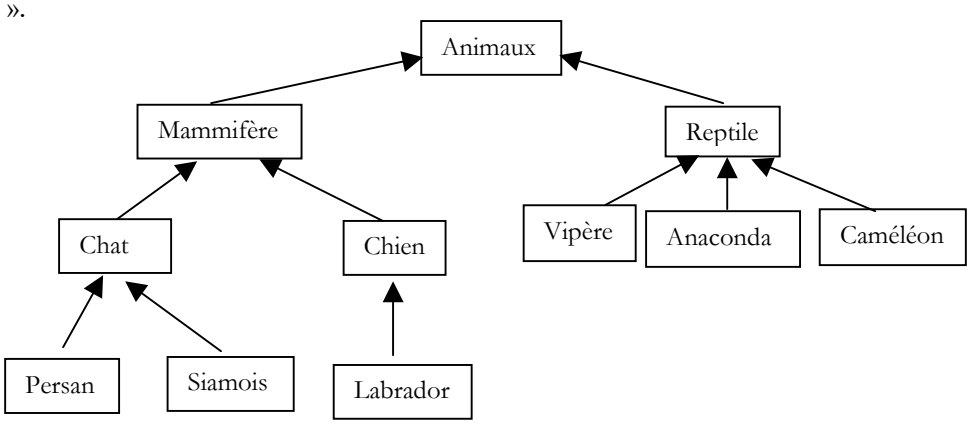


Figure 3.1 : Exemple de taxonomie

2.1.1 Mesures reposant sur la distance

Les mesures reposant sur la distance considèrent que la similarité entre deux concepts peut être calculée à partir du nombre de liens qui séparent les deux concepts. Plusieurs variantes existent en fonction du chemin pris en compte pour calculer la distance entre les concepts.

La mesure du **edge counting** [Rada 1989] évalue la distance sémantique à partir du nombre de branches séparant les concepts par le plus court chemin dans la hiérarchie.

A partir de l'exemple de la figure 3.1, la distance se calcule de la façon suivante :

$Dist_{edge}(siamois, persan)=2$

$Dist_{edge}(reptile, anaconda)=1$ $Dist_{edge}(chat, mammifère)=1$

$Dist_{edge}(persan, reptile)=4$ $Dist_{edge}(persan, labrador)=4$

La similarité sémantique entre deux concepts correspond à l'inverse de la distance entre deux concepts. Plus deux concepts sont distants, moins ils sont similaires. A partir de la mesure de distance précédente, [Leacock 1998] a proposé la formule suivante pour calculer la similarité. Elle est issue de la proposition de [Resnik 1998] et assure la normalisation de cette dernière.

$$\mathrm{Sim}_{\mathrm{edge}}(c1,c2) = -\log\frac{Dist_{edge}(c1,c2)}{2*Max}$$

Où max étant la profondeur maximale de la taxonomie

L'utilisation de la fonction –log permet de normaliser la similarité entre [0,1] (1 signifiant que les concepts sont totalement similaires).

sim_{edge} *(siamois, persan)= -log* $\frac{2}{2*4}$ *=0,6*

sim_{edge} *(reptile, anaconda)= 0,9* sim_{edge} *(chat, mammifère)=0,9*

sim_{edge} *(persan, reptile)=0,3* sim_{edge} *(persan, labrador)=0,3*

Wu et Palmer [Wu 1994] ont proposé une autre mesure de similarité prenant en compte à la fois la profondeur des concepts dans la hiérarchie de concepts et la structure de la hiérarchie de concepts.

Pour calculer la similarité entre deux concepts c1 et c2, la formule suivante est utilisée :

$$Sim_{Wu}(c1,c2)=\frac{2*depth(c)}{depth(c1)+depth(c2)}$$

où depth(c_i) correspond au niveau de profondeur du concept c_i
et c représente le concept le plus spécifique qui généralise c1 et c2.

La valeur de la similarité est comprise entre 0 et 1 (1 signifiant que les concepts sont totalement similaires).

sim_{Wu} *(siamois, persan)= 2*3/(4+4)=3/4 (chat étant le plus spécifique subsumeur)*

sim_{Wu} *(reptile, anaconda)= 2*2/(2+3)=4/5* sim_{Wu} *(chat, mammifère)=4/5*

sim_{Wu} *(persan, reptile)=2*1/(2+4)=1/3* sim_{Wu} *(persan, labrador)=2*2/(4+4)=1/2*

Cette mesure est plus pertinente que les mesures précédentes reposant uniquement sur le chemin le plus court entre les deux concepts, car elle prend en compte l'organisation hiérarchique des concepts, c'est-à-dire le concept généralisant les deux concepts considérés. La similarité entre *persan* et *reptile* devient plus basse par sim_{Wu} que celle entre *persan* et *labrador*, alors qu'elles étaient identiques par les mesures reposant sur le « edge counting ». Cette différence est pertinente dans la mesure où *labrador* et *persan* ont un concept commun les spécifiant *(mammifère)* plus proche ces deux concepts dans la hiérarchie que *persan* et *reptile*. Cependant, cette mesure admet que la distance sémantique entre deux concepts reliés par la relation « est un » est égale, alors que ce n'est pas forcément le cas. La distance sémantique portée par le lien « est un » entre *reptile* et *anaconda* et celle portée par ce même type de lien entre *mammifère* et *chat* est considérée de la même façon alors que ces deux relations ne témoignent pas du même niveau de spécificité (on passe directement de la catégorie *reptile* à l'espèce *anaconda* sans spécifier les catégories intermédiaires comme cela était le cas pour la catégorie *mammifère*). Les choix arbitraires pris lors de la construction de la hiérarchie de concepts influencent donc la valeur de la similarité. Afin de prendre en compte le fait que les liens dans une ontologie ne représentent pas la même distance sémantique, plusieurs solutions ont été proposées.

L'une d'elles consiste à prendre en compte la densité locale de l'ontologie au niveau des concepts considérés pour contrecarrer la subjectivité dans le choix des relations. L'utilisation de la densité repose sur l'observation suivant laquelle les concepts appartenant à une partie dense en concepts de l'ontologie sont sémantiquement plus proches que ceux appartenant à une partie éparse de l'ontologie [McHale 1998]. Dans [Richardson 1995] la densité est représentée par le nombre de liens fils du concept. Plus un concept a de fils, plus la similarité entre le concept et ses fils est élevée. Le point faible de ce critère est que la distribution des concepts dans l'ontologie doit refléter la distribution des concepts dans le domaine. Cependant, comme nous l'avons vu précédemment, les choix faits lors de l'élaboration de la hiérarchie de concepts sont généralement

liés à la connaissance utile pour la tâche pour laquelle elle est construite et ne reflète pas forcément le domaine dans son ensemble.

Une autre solution consiste à prendre en compte le contenu en information des concepts.

2.1.2 Mesures reposant sur le contenu en information des concepts

Les approches reposant sur le contenu en information supposent que l'information détenue par les concepts puisse être quantifiée. La similarité est alors calculée à partir de l'information partagée par les concepts. Les différentes mesures proposées pour évaluer la similarité entre concepts reposent sur ce calcul.

2.1.2.1 Calcul du contenu en information d'un concept

Afin de calculer l'information contenue par les concepts, un corpus de référence est choisi. Les concepts sont alors pondérés par une fonction correspondant à l'information portée par un concept dans le corpus. Le contenu en information du concept est défini par ses occurrences dans le corpus ainsi que celles des concepts qu'il subsume. Il a pour objectif d'utiliser la probabilité d'obtenir un concept dans un corpus documentaire afin de contrecarrer la subjectivité dans le choix des relations « est un » de l'ontologie.

Le contenu en information d'un concept c se calcule de la façon suivante [Resnik 1995] :

$$CI(c) = -\log(p(c))$$

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

où word(c) est l'ensemble des termes ou labels représentant le concept c et les concepts subsumés par c, count(n) le nombre d'occurrences du terme n dans le corpus et N le nombre total d'occurrences des labels de concepts retrouvés dans le corpus. L'utilisation de la fonction – log permet de réduire le contenu en information d'un concept ayant une forte probabilité d'apparition dans le corpus.

Un inconvénient du calcul du contenu en information d'un concept proposé par Resnik est que la probabilité d'obtenir un concept est calculée à partir des labels retrouvés dans le corpus sans vérifier que chacune des occurrences se rapporte effectivement au concept. Ceci peut poser problème dans le cas où le label est ambigu et désigne différents concepts. Le contenu en information d'un concept peut ainsi être amplifié par la forte occurrence d'un label désignant un autre concept dans le corpus.

Une solution plus adaptée serait de désambiguiser le label dans les documents et de comptabiliser uniquement les occurrences qui se rapportent au concept en question. Cette solution ne peut pas être mise en place dans la mesure où [Resnik 1999] utilise ces mesures justement pour désambiguïser les termes. Afin de prendre en compte les labels ambigus, une autre formule a été proposée dans [Sanderson 1995]. Dans cette formule (freq2), la fréquence d'apparition d'un label est normalisée par rapport au nombre de concepts auxquels il se rapporte.

$$Freq2(c) = \sum_{n \in word(c)} \frac{count(n)}{nbclasse(n)}$$

où nbclasse(n) correspond au nombre de concepts dont le terme n est label

2.1.2.2 Interprétation du contenu en information

Plusieurs mesures de similarité ont été définies à partir du contenu en information des concepts [Resnik 1995] [Lin 1998] [Jiang 1997]. Elles reposent sur le principe que l'information

commune aux deux concepts est capturée dans le contenu en information du concept le plus spécifique qui subsume les deux concepts.

Afin d'illustrer l'intérêt de la prise en compte du contenu en information des concepts, la figure 3.2 reprend la hiérarchie de concepts présentée dans la figure 3.1 pour laquelle le contenu en information de chaque concept est précisé.

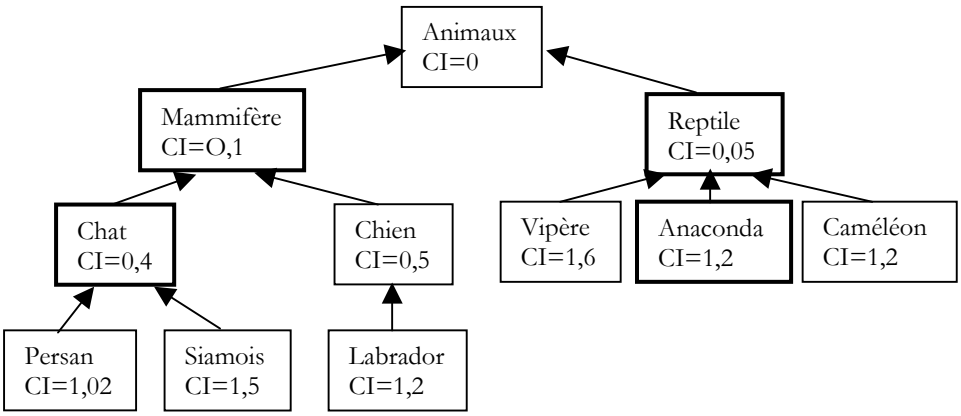


Figure 3.2 Hiérarchie de concepts augmentée par le contenu en information (CI) des concepts.

En prenant par exemple les concepts *Chat* et *Mammifère*, l'information partagée par ces deux concepts est le contenu en information du concept *Mammifère*, c'est-à-dire 0,1. L'information commune aux concepts *Reptile* et *Anaconda* est 0,05. Ainsi l'information partagée par deux concepts permet de capturer le poids sémantique du lien « est un ».

Afin de pouvoir calculer la similarité entre deux concepts, la probabilité d'obtenir dans le corpus le concept le plus spécifique subsumant c1 et c2 est définie de la façon suivante :

$$\mathrm{Pms}(c1,c2) = \min_{c \in S(c1,c2)} \{p(c)\}$$

où S(c1,c2) est l'ensemble de concepts qui subsument à la fois c1 et c2

La probabilité d'obtenir un concept prend en considération à la fois la probabilité d'obtenir un concept ainsi que tous les concepts qu'il subsume. Pms (ou min{p(c)}) revient donc à choisir le concept subsumant c1 et c2 ayant la plus faible probabilité, c'est-à-dire le concept le plus spécifique de c1 et c2 dans la hiérarchie au sens où son contenu en information est le plus faible. Quand l'ontologie permet l'héritage multiple et que plusieurs concepts subsument les deux concepts considérés, cette formule permet de choisir dans l'ensemble des concepts candidats le concept le plus spécifique au sens où c'est celui qui a la probabilité la plus faible.

2.1.2.3 Mesures

La première mesure de similarité calculée à partir du contenu en information des concepts a été proposée dans [Resnik 1995]. Elle prend des valeurs entre [0,1].

$$\mathrm{Sim}_{\mathrm{Resnik}}(c1,c2) = [-\log(Pms(c1,c2)]$$

Sim_{Resnik} (siamois, persan)= CI(chat)=0,4 (chat étant le plus spécifique subsumeur)

sim_{Resnik} (reptile, anaconda)= 0,05 sim_{Wu} (chat, mammifère)=0,1

sim_{Resnik} *(persan, reptile)=0* sim_{Wu} *(persan, labrador)=0,1*

L'inconvénient de cette mesure est que deux couples de concepts qui ont le même subsumeur le plus spécifique ont la même similarité. Ceci est par exemple le cas entre (*persan* et *labrador)* et (*chat* et *mammifère*).

La mesure proposée par Lin vise à contrecarrer cet inconvénient [Lin 1998]. Dans le cas de l'évaluation de la similarité entre deux concepts dans une taxonomie, Lin étend ces mesures en considérant que la description de chacun de ces concepts est le contenu en information des concepts et que les caractéristiques communes aux deux concepts sont quantifiées par le contenu en information du concept le plus spécifique généralisant les deux concepts.

$$\mathrm{sim}_{\mathrm{Lin}}(c1,c2)=\frac{2*\log(pms(c1,c2))}{\log(p(c1))+\log(p(c2))}$$

Cette mesure a l'avantage de prendre en compte le concept le plus spécifique subsumant c1 et c2 ainsi que le contenu en information des concepts comparés. Contrairement à la mesure précédente proposée par Resnik, elle permet de différencier la similarité entre plusieurs couples de concepts ayant le même subsumeur le plus spécifique. Elle prend des valeurs entre [0,1].

Sim_{Lin} *(siamois, persan)=* $\frac{2*CI(chat)}{CI(siamois)+CI(persan)}$ *=0,32*

Sim_{Lin} *(reptile, anaconda)= 0,08* sim_{Lin} *(chat, mammifère)=0,4*

Sim_{Lin} *(persan, reptile)=0* sim_{Lin} *(persan, labrador)=0,09*

2.1.3 Mesures Mixtes

Le principe des mesures mixtes est de considérer le plus court chemin reliant deux concepts dans l'ontologie et de pondérer ces liens à partir de leur poids sémantique. Le poids sémantique des liens prend notamment en compte le contenu en information des concepts [Richarson 1995] [Jiang 1997]. Aucune formule exacte n'étant donnée dans [Richarson 1995], nous choisissons pour présenter ce type d'approche celle présentée dans [Jiang 1997].

Les liens du plus court chemin reliant les deux concepts sont pondérés à partir de quatre facteurs.

- Le contenu en information du lien calculé à partir de la différence du contenu en information du concept fils c_{fils} et celui du père $c_{père}$

$$CI(c_{fils},c_{père})=CI(c_{fils})-CI(C_{père})$$

- La profondeur du concept père $c_{père}$ évaluée de la façon suivante :

$$\mathrm{Prof}(c_{père})=\frac{profondeur(c_{père})+1}{profondeur\_maxi}$$

- La densité locale du concept père calculée à partir du nombre de nœuds fils du concept $E(c_{père})$ et le nombre moyen $\overline{E}$ de nœud fils pour un concept dans le réseau.

$$\mathrm{Densité}(c)=\frac{\overline{E}}{E(c)}$$

- Le poids accordé au type de lien considéré $t(c_{fils}, \overline{E})$.

Le poids d'un lien est ensuite calculé de la façon suivante où α et β permettent de pondérer les différents facteurs avec $\alpha \geq 0$ et $0 \leq \beta \leq 1$:

$$Poids(c_{fils}, c_{père}) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(c père)}\right)\left(\frac{d(c père)+1}{d(c père)}\right)^{\alpha} [CI(C_{fils}) - CI(C_{père})]T(c_{fils}, c_{père})$$

La distance entre deux concepts c1 et c2 est alors calculée par :

$$Dist_{Jiang}(c1,c2) = \sum_{c \in pcc(c1,c2)} Poids(c, père(c))$$

où pcc(c1,c2) représente l'ensemble des concepts appartenant au plus court chemin entre c1 et c2.

Dans le cas où seul le contenu en information des liens est considéré (α et β sont nuls et le poids du lien est 1), la distance entre deux concepts peut être simplifiée par la formule suivante :

$$Dist_{Jiang}(c1,c2) = CI(c1) + CI(c2) - 2*CI(SPS(c1,c2))$$

où SPS(c1,c2) représente le concept le plus spécifique subsumant c1 et c2

Cette dernière formule est équivalente à celle proposée par Lin. La seule différence est qu'elle permet de calculer la distance sémantique et non pas la similarité. Jiang propose de convertir la distance proposée en similarité en adaptant la formule de similarité proposée par Resnik.

$$Sim_{Jiang}(c1,c2) = (2*max) - Dist_{jiang}(c1,c2)$$

où max représente la distance maximale obtenu par la formule $Dist_{Jiang}$

2.1.4 Evaluation des mesures

Les différentes mesures présentées dans cette section ont été évaluées et comparées dans plusieurs travaux [Resnik 1995] [Jiang 1997] [Lin 1998] [Budanitsky 2001]. L'évaluation consiste à comparer la valeur donnée par les mesures pour différentes paires de termes avec des valeurs de similarités affectées par des humains. Un ensemble de 30 paires de référence a été confectionné par Miller et Charles [Miller 1991] d'après le jugement de 38 étudiants. Chacun des étudiants a affecté à ces paires un score de similarité croissant allant de 0 à 4 (0 correspondant à non sémantiquement lié, et 4 à fortement synonyme). La moyenne des similarités permet d'estimer le degré de similarité obtenues par un jugement humain. Les résultats obtenus par deux études [Miller 1991] [Resnik 1995] sont très proches (corrélation de 0,96), indiquant que les estimations sont effectivement représentatives d'un jugement humain.

Les similarités affectées par des humains et celles affectées par les mesures ont ensuite été corrélées. Les termes de l'ensemble d'évaluation étant du domaine de langue générale, les mesures ont été testées en utilisant la ressource WordNet [Miller 1988] et le thésaurus Roget [Roget 1962]. Pour les mesures reposant sur le contenu en information, le corpus utilisé pour calculer le contenu en information des concepts est soit le corpus Brown dans son intégralité [Resnik 1995], soit un sous-ensemble de ce corpus appelé SemCor [Miller 1993] dont chacun des termes a été étiqueté à partir des synsets de WordNet [Lin 1998] [Jiang 1997]. L'avantage de ce dernier est qu'il permet de calculer le contenu en information à partir de la fréquence d'apparition des termes qui réfèrent effectivement le concept (cf notre remarque sur la limite du calcul de fréquence proposé par Resnik). Afin de pouvoir comparer les différentes approches reposant sur le contenu en information, les résultats que nous considérons sont ceux du corpus SemCor. Une comparaison des différentes mesures sur l'ensemble des paires a été menée dans différents

travaux [Lin 1998] [Jiang 1997] et [Mc Hale 1998]. En considérant la hiérarchie de concepts issue de Wordnet, la mesure qui permet d'obtenir les résultats les plus proches des jugements humains est la méthode proposée par Lin, suivie de très près la mesure de Jing. La pertinence de l'utilisation du contenu en information des concepts est alors vérifiée. Cette conclusion n'est pas vraie pour les évaluations utilisant le thésaurus Roget. Pour ce thésaurus, les meilleurs résultats sont obtenus par la mesure reposant sur le « edge couting ». Cette différence s'explique par le fait que le corpus considéré pour le calcul du contenu en information des concepts est particulièrement adapté à l'utilisation de WordNet car le corpus est désambiguïsé à partir de ses synsets.

Des limites à ce type d'évaluation ont cependant été mises en évidence dans [Budanisky 2001]. Les étudiants choisis pour réaliser l'expérimentation devaient déterminer la similarité des paires de mots en choisissant le sens dominant des termes, c'est-à-dire le concept le plus susceptible d'être référencé par les termes. Cependant, les mesures ne visent pas à évaluer la similarité entre les termes mais entre le sens des concepts. L'évaluation devrait porter sur les concepts directement, pour éviter le biais introduit par le choix du sens déterminé par les étudiants. La connaissance de chacun des étudiants n'est pas non plus prise en compte, leur jugement a priori peut également biaiser l'évaluation.

2.2 Similarité dans une ontologie faisant intervenir des liens associatifs

Plusieurs travaux visent à permettre le calcul de la similarité entre concepts en ne limitant pas les liens considérés aux liens taxonomiques.

Dans [Lord 2003], les mesures de Resnik, Lin et Jiang sont utilisées pour calculer la similarité entre concepts à partir des liens « est un » et « partie de » de l'ontologie des Gènes. Les mesures sont utilisées sans modification, en considérant les deux relations comme étant la relation père-fils permettant d'organiser hiérarchiquement les concepts. Le contenu en information d'un concept est calculé à partir de la probabilité d'obtenir les labels du concept dans le corpus ainsi que chacun de ses fils auxquels il est lié aussi bien par la relation « est un » que par la relation « partie de ». Cette considération a, selon nous, des limites. Dans ces mesures, le contenu en information du concept le plus spécifique lié aux deux concepts est considéré pour évaluer l'information commune à deux concepts. Cependant, nous considérons que le subsumeur de deux concepts ne peut être calculé qu'à partir du moment où une relation sémantique peut être établie entre celui-ci et les deux concepts considérés, ceci impliquant que le concept subsumeur contient effectivement des informations sur les deux concepts. Prenons l'exemple présenté dans la figure 3.3, il paraît intuitif que l'information commune à *Part* et *Pique-nique* n'est pas détenue par *Repas* qui généralise effectivement *pique-nique* mais pas *part*.

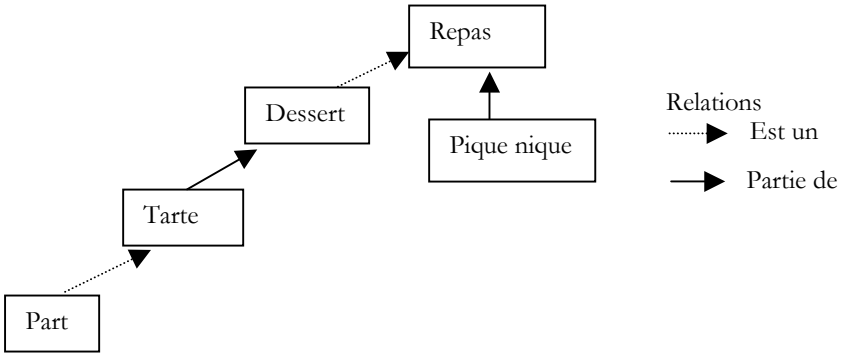


Figure 3.3 Exemple d'ontologie faisant intervenir deux relations « partie de » de familles différentes

L'adaptation des mesures reposant sur le contenu en information des concepts doit considérer le fait qu'une relation sémantique puisse être trouvée entre les deux concepts dont la similarité est calculée.

Dans [Thieu 2004], une adaptation de la mesure de Jiang est proposée pour prendre en compte les relations non taxonomiques. La distance sémantique impliquée par les relations taxonomiques est considérée de la même façon que dans la méthode initiale. L'adaptation porte uniquement sur le poids sémantique affecté au lien non taxonomique. Jiang propose de représenter le poids à partir de la différence du contenu en information du fils et du père. Or, dans le cas des relations non taxonomiques, le contenu en information n'est pas décroissant du père au fils car le père ne généralise pas le concept fils. De plus, le contenu en information des deux concepts calculé par la probabilité d'obtenir les labels des concepts et ceux de leur subsumeur ne contient aucun élément commun. Les auteurs proposent alors de pondérer le lien au moyen des différences de contenu en information du fils avec l'ensemble de ses pères taxonomiques. Bien que cette adaptation ait l'avantage de s'adapter à n'importe quel type de lien, elle ne prend pas en compte le contenu en information du nœud père et ceci implique une perte d'information. Le poids de tous les liens non-taxonomiques partant d'un même neoud est en effet égal.

D'autres mesures ne prenant pas en compte le contenu en information des concepts ont été proposées.

Dans [Bernstein 2005], une mesure reposant sur la représentation vectorielle des concepts est présentée. L'espace de représentation des concepts est l'ensemble des relations qui peuvent être liées à un concept. Ces relations sont soit des relations d'attribut, soit des relations liant le concept à d'autres concepts dans l'ontologie. Par exemple, un concept *chaise* ayant quatre pieds et étant lié au concept *dossier* par la relation *a comme partie* et au concept *bureau* par la relation *est situé dans* est représenté dans l'espace vectoriel [« *a comme partie pied* », « *a comme partie dossier* », « *est situé dans bureau* »] par le vecteur [4,1,1]. Bien qu'elle soit efficace d'un point de vue computationnel, cette approche a pour inconvénient principal le fait que le lien sémantique entre les relations constituant l'espace vectoriel ne soit pas considéré.

Dans [Ehrig 2005], la mesure de similarité reposant sur le nombre d'arcs taxonomiques du plus court chemin entre les concepts, proposée par Rada, est utilisée pour calculer la similarité sémantique à partir de n'importe quel type de relation composant le plus court chemin. L'inconvénient de cette approche est d'admettre que les relations sémantiques représentent la même distance sémantique entre les concepts. Ceci pose problème parce que chaque type de relations implique sa propre sémantique. Aussi, comme nous l'avons vu précédemment, une même relation, telle que la relation « est un », peut impliquer un degré d'engagement différent.

2.3 Bilan

La plupart des mesures de la littérature ne considèrent que les liens taxonomiques. Pourtant la connaissance modélisée par les autres liens ajoute de la sémantique aux concepts. Les mesures qui incluent tous les types de liens ont l'inconvénient soit de ne pas différencier le poids des liens en fonction de leur type, soit d'ignorer une partie de la sémantique qu'ils représentent (elles ne considèrent qu'un des deux concepts de la relation ou aucun lien entre les relations).

Les mesures que nous considérons dans nos travaux sont celles reposant sur le contenu en information des concepts car ces mesures permettent de contrecarrer la subjectivé dans le choix des relations lors de la construction de l'ontologie. Les évaluations de la littérature valident ce choix. Nous proposons de les étendre pour prendre en compte d'autres types de relations non taxonomiques. D'autre part, l'évaluation de cette nouvelle mesure (chapitre 7) n'a pas les limites

des évaluations de la littérature puisque les couples donnés à évaluer aux experts sont composés de concepts et non pas de termes. Le contexte de chaque concept dans l'ontologie (labels, relations associatives et taxonomiques) est précisé afin que l'expert cerne le sens associé au concept. De plus, nous ne faisons pas appel à des étudiants mais à des experts du domaine.

3 Quelle ontologie choisir ?

La majorité des approches de RI visant à intégrer une ontologie dans leur procédé reposent sur des ontologies existantes [Hearst 1997][Vallet 2005][Baziz 2005]. Généralement, l'unique caractéristique prise en compte dans le choix de l'ontologie est le domaine de connaissance représentée dans l'ontologie qui doit couvrir le domaine traité dans le corpus. C'est le cas par exemple du système Cat-a-cone qui repose sur la hiérarchie de concepts du domaine de la médecine MESH [Hearst 1997] pour explorer une collection documentaire du même domaine, ou bien des travaux présentés dans [Baziz 2005] qui repose sur l'ontologie générale WordNet pour une tâche de RI ad-hoc sur une collection de TREC. Cependant, nous pensons que ce choix n'est pas aussi simple. Les ontologies considérées doivent être adaptées aux tâches de RI visées et surtout apporter de la connaissance pertinente par rapport à l'information présente dans les corpus. Ceci est également l'explication donnée dans de nombreux travaux pour justifier le fait que l'indexation sémantique ait besoin d'être couplée à une indexation classique afin d'améliorer les performances des systèmes [Vallet 2005] [Mihalcea 2000].

L'objectif de cette section est donc d'analyser dans quelle mesure les ontologies sont réutilisables dans des perspectives de RI. Pour cela, nous verrons dans un premier temps les positions prises par les concepteurs d'ontologies par rapport à leur réutilisabilité. Puis nous analyserons des travaux portant sur différents types d'analyses permettant d'évaluer la réutilisabilité des ontologies. Finalement, nous présenterons plusieurs approches portant sur l'évaluation de l'adéquation de hiérarchies de concepts à des corpus. Ces dernières approches ont pour but de choisir une représentation de la connaissance se rapportant au mieux aux contenus des granules documentaires.

3.1 Réutilisabilité des ontologies

La construction d'ontologies réutilisables est le but affiché d'un certain nombre de travaux [Gomez-Perez 1996] [Fernandez 1997] [Uschold 1996].

Cependant, de nombreux auteurs considèrent que les ontologies sont non réutilisables. Bachimont affirme en effet que par leur méthode de construction et les travaux epistémologiques qui les supportent, leur réutilisation est impossible [Bachimont 1996]. De la même façon, Charlet considère que « les ontologies sont des artefacts construits en fonction d'une tâche précise et ne peuvent être réutilisées, en tant qu'objet formel, pour une autre tâche. ». Il affirme également que les travaux sur la génération d'ontologies à partir de corpus montrent une dépendance forte entre la construction de corpus et la construction de la future ontologie. « Le corpus est porteur via les expressions linguistiques qui en sont extraites, des futurs concepts de l'ontologie » [Charlet 2002]. Une ontologie construite à partir d'un corpus donné ne serait donc pas adaptée à un nouveau corpus. Nous partageons le point de vue de Furst [Furst 2004] selon lequel les ontologies sont destinées à être réutilisées. La sémantique qu'elles représentent est liée au cadre applicatif à partir duquel le sens des termes et concepts est défini. Cependant, la représentation ne dépend pas de l'opération faite avec l'ontologie. La sémantique de l'ontologie est liée au contexte mais la représentation n'implique pas que l'ontologie soit utilisée uniquement dans le contexte de sa création.

3.2 Evaluer la réutilisation d'une ontologie

Afin d'évaluer la réutilisabilité des ontologies plusieurs démarches sont suivies.

La première consiste à considérer une ontologie existante et à décrire les étapes et le coût impliqués par le processus de réutilisation dans une application donnée. Cette démarche est suivie notamment par Uschold qui recommande la création d'ontologies à partir de la réutilisation d'ontologies existantes plutôt qu'en partant de rien [Uschold 1995]. Les travaux présentés dans [Uschold 1998] et [Pinto 2001] proposent une analyse du procédé impliqué par la réutilisation d'une ontologie formelle pour la construction d'une nouvelle et les étapes nécessaires à l'application de cette ontologie dans un nouveau système. Les conclusions de ces deux analyses montrent que l'automatisation de ce procédé est loin d'être envisageable dans la mesure où il nécessite des connaissances extérieures liées à l'ontologie et à la tâche à réaliser.

Une autre approche consiste à évaluer la réutilisabilité de l'ontologie par rapport à certains critères voire certaines mesures.

Une première tentative consiste à reprendre les mesures d'évaluation des SRI basées sur les notions de précision (proportion de documents correctement retournés par rapport à l'ensemble des documents retournés par le système) et de rappel (proportion de documents correctement retournés par rapport à l'ensemble des documents pertinents dans la collection) [Salton 1971]. Cependant, l'évaluation d'une ontologie n'est pas aussi triviale [Brewtser 2004]. Les notions de rappel et de précision devraient être comprises de la façon suivante. La précision correspondrait à évaluer la quantité de connaissance correctement identifiée dans l'ontologie par rapport à toute la connaissance contenue dans l'ontologie en fonction de la tâche à réaliser. Le rappel serait la quantité de connaissance correctement définie dans l'ontologie par rapport à la connaissance qui devrait être identifiée. Le problème est qu'il est impossible de déterminer ces ensembles de connaissances ; ils dépendent en effet des différentes interprétations et des différents types de connaissance que l'on souhaite représenter.

D'autres solutions ont donc été proposées pour permettre l'évaluation d'une ontologie. Elles peuvent être regroupées en deux types d'analyse: l'analyse qualitative ou l'analyse quantitative (section 3.2.1). Ces analyses peuvent être appliquées pour évaluer l'adéquation entre une hiérarchie de concepts et un corpus (section 3.2.2).

3.2.1 Analyse qualitative et analyse quantitative

Une **analyse qualitative** consiste à évaluer une ontologie ou ses parties et à mesurer son taux de pertinence. Cependant, se posent deux questions : qui parmi l'utilisateur de l'ontologie, un ou plusieurs experts du domaine et le concepteur de l'ontologie est censé donner ce taux ? quels sont les critères qui devront être pris en compte dans l'évaluation ?

[Guarino 1998b] et [Gomez Perez 1999] répondent à cette dernière question en proposant des critères fondés sur les principes utilisés lors de la construction de l'ontologie. Les critères proposés par [Gomez Perez 1999] sont les suivants :

- la consistance de l'ontologie : la possibilité d'obtenir des conclusions contradictoires à partir des inférences possibles sur l'ontologie est ici évaluée,
- la complétude de l'ontologie : l'ontologie recouvre toute la connaissance qu'elle est censée représenter et chacune de ses définitions contient bien tous les éléments nécessaires,
- la concision de l'ontologie : l'ontologie ne contient pas de connaissance inutile ou redondante,
- l'expansibilité de l'ontologie : l'ajout de connaissance dans l'ontologie est possible,
- la sensibilité de l'ontologie : le changement d'une définition n'altère pas toutes les autres définitions.

Ces critères restent cependant très théoriques et nécessitent leur évaluation par les concepteurs de l'ontologie, ceux-ci étant capables de les évaluer à partir de la sémantique qu'ils associent à ces critères. Ces critères sont de plus indépendants de la tâche pour laquelle l'ontologie est réutilisée. Ce type de critères peut également amener à la construction d'ontologies non opérationnelles par l'absence de la prise en compte de la tâche [Milks 2002].

[Lozano-Tello 2004] propose de nouveaux critères en créant un cadre d'évaluation multidimensionnelle aidant au choix d'une ontologie pour une tâche donnée. Les critères pris en compte sont regroupés sous forme hiérarchique à partir de cinq dimensions : le contenu de l'ontologie, le langage de représentation, la méthodologie suivie lors de la construction, les outils pouvant utiliser l'ontologie, et le coût de la réutilisation. Chaque dimension est ensuite déployée par rapport à des caractéristiques, pouvant elles-mêmes être détaillées par des sous caractéristiques plus spécifiques. L'ensemble de ces critères est pondéré par un expert en développement d'ontologie en fonction du cadre de réutilisation de l'ontologie. Une analyse multidimensionnelle des valeurs affectées aux différents critères permet ensuite à l'expert de choisir quelle ontologie utiliser. L'inconvénient majeur de cette approche est qu'elle nécessite un investissement important de la part de l'expert dans l'évaluation des critères.

Une **analyse quantitative** consiste quant à elle à évaluer la réutilisabilité d'une ontologie par rapport à son efficacité dans la réalisation d'une tâche donnée. Une évaluation de ce type consisterait par exemple à prendre plusieurs ontologies différentes et à exécuter une même tâche avec chacune d'entre elles, puis à comparer les résultats obtenus. Dans la mesure où une ontologie doit permettre l'interopérabilité entre applications, il est possible de concevoir son évaluation par rapport aux résultats qu'elle permet et non pas par rapport à l'interprétation qu'en fait un être humain [Brewster 2004]. Un exemple d'analyse quantitative est donné dans [Porzel 2004]. L'évaluation repose sur la définition d'une tâche, la considération d'une ou plusieurs ontologies dites légères (au sens où nous l'avons défini dans le chapitre 1), la définition d'une application réalisant la tâche par l'intermédiaire d'un algorithme prenant en compte l'ontologie, et la définition de références donnant les réponses correctes qui doivent être retournées par l'application. L'efficacité de l'ontologie est examinée par rapport aux différents niveaux de l'ontologie (vocabulaire, taxonomie, relation non taxonomique) suivant trois critères : les erreurs d'insertion (concepts, relations taxonomiques, relations non taxonomiques non appropriés), les erreurs de délétion (concepts, relations taxonomiques, relations non taxonomiques manquants) et les erreurs de substitution (concepts, relations taxonomiques, relations non taxonomiques ambigus). La tâche choisie est l'étiquetage par des relations ontologiques des relations entre entités de l'ontologie trouvées dans des textes. Cette tâche peut être apparentée à un système de désambiguïsation dont le but est d'identifier les concepts abordés dans le texte et d'extraire les relations sémantiques liant ces concepts. Le point critique de cette approche est la réalisation de l'ensemble de références. Elle demande l'intervention d'experts et pénalise l'évaluation car elle implique qu'une personne réalise la tâche.

Dans [Maedche 2002b], des mesures sont définies pour comparer deux ontologies entre elles et plus particulièrement un ensemble d'ontologies par rapport à une ontologie de référence. Les ontologies considérées dans ces travaux sont des ontologies dites légères. La comparaison des deux ontologies repose sur une analyse lexicale (comparaison des termes des ontologies) et une analyse conceptuelle (comparaison de l'organisation des concepts dans l'ontologie). Ces mesures donnent des résultats concluants dans la comparaison d'une ontologie de référence d'un domaine réalisée par des experts et des ontologies du même domaine réalisées par des étudiants. Cependant, dans des perspectives de RI, l'objectif n'est pas de comparer des ontologies entre elles mais d'évaluer la ou les plus pertinentes pour la tâche de RI. Dans ce cadre là, la comparaison doit prendre en compte le corpus sur lequel la tâche doit être réalisée parce que l'ontologie est utilisée pour aider au traitement du corpus.

Un autre type d'analyse quantitative consiste justement à comparer l'adéquation entre une ou plusieurs ontologies par rapport à un corpus. Ce type d'analyse permet d'évaluer si la connaissance contenue dans l'ontologie se rapporte à l'information contenue dans un corpus. Dans des perspectives de RI, ce type d'analyse est pertinent dans la mesure où il permet de choisir une ontologie qui permettra aux SRI de s'appuyer sur de la connaissance utile pour la compréhension et l'interprétation du corpus. Différents travaux portant sur cette thématique ont pour but de comparer un corpus avec des hiérarchies de concepts. Ils sont décrits dans les sous sections suivantes.

3.2.2 Adéquation d'une hiérarchie de concepts à un corpus

Desmontils et Jacquin [Desmontils 2002] proposent d'évaluer la pertinence d'une ontologie légère par rapport aux sites Web à indexer. Seule la relation « est un » est considérée dans l'analyse.

Dans le but d'évaluer l'adéquation, différents facteurs sont définis.

- le nombre de concepts directement retrouvés dans les pages des sites : DID,
- le nombre de concepts indirectement retrouvés dans les pages (nombre de concepts subsumant les concepts retrouvés) : IID
- le nombre de pages ayant un concept de l'ontologie : OCD
- le pouvoir représentatif des concepts retenus pour représenter les pages : MCR. Le pouvoir représentatif d'un concept prend en compte la fréquence d'apparition des termes désignant le concept dans chacune des pages mais également ses relations avec les autres concepts dans une ontologie de référence retrouvés dans la page. Plus un concept a des relations avec les autres concepts présents dans la page, plus il est représentatif de la page. La ressource conceptuelle utilisée comme référence est WordNet.

Afin d'évaluer la pertinence d'une ontologie O pour un corpus S, les différents facteurs sont combinés linéairement et définissent le facteur d'adéquation de l'ontologie au site Web.

$$OSAD_{S,O} = \frac{IID_{S,O}}{2} + DID_{S,O} + 2*OCD_{S,O} + 2*MRC_{S,O}$$

L'influence de ces différents facteurs a été évaluée en calculant l'adéquation entre l'ontologie SHOE et le site Web de l'université de Washington. Le facteur OSAD n'est que de 56% car l'ontologie ne couvre pas tous les domaines d'étude. Des conclusions ont été tirées sur l'impact des différents facteurs.

- Les facteurs OCD et MRC, lorsqu'ils sont bas, indiquent que certaines pages ne pourront pas être indexées,
- Un facteur DID bas implique la nécessaire mise à jour de l'ontologie,
- Si tous les coefficients et donc OSAD sont bas, une autre ontologie doit être choisie.

La limite de cette approche est qu'elle repose sur une ressource extérieure à l'ontologie choisie, ici WordNet, pour indexer le corpus. La représentativité d'un concept devrait être calculée au sein de l'ontologie choisie et non pas à partir de cette ressource. De plus, seules les relations taxonomiques sont considérées, ce qui limite le niveau de sémantique considéré. Comme expliqué dans la section 1, la mesure de similarité choisie omet également certaines considérations sémantiques dans la conception des ressources.

Rottenburger [Rottenburger 2002] propose un autre cadre d'évaluation d'ontologies par rapport à un corpus. Dans cette approche, les ontologies sont également ramenées à des hiérarchies de concepts. La principale différence avec l'approche précédente est qu'elle n'a pas pour objectif de déterminer les avantages et les lacunes de l'ontologie pour la représentation du corpus mais de définir des indicateurs permettant de comparer la connaissance qu'elle représente par rapport à différents corpus. Le pouvoir représentatif d'un concept n'est plus évalué uniquement par rapport à sa représentativité dans chacun des documents mais également par rapport à sa représentativité dans l'ensemble du corpus. Cette considération est déterminante car elle permet de s'assurer que l'ontologie spécifie effectivement le contexte associé au corpus et que les concepts de chacun des documents sont liés entre eux. Un premier ensemble d'indicateurs d'adéquation entre le corpus S et l'ontologie O a été défini. Les indicateurs suivants ne prennent pas en compte la représentativité des concepts dans les documents :

- le nombre de concepts de la hiérarchie de concepts de O retrouvés dans le corpus S (DID).
- le nombre de termes du corpus S retrouvés dans O : TC
- la couverture de O sur le corpus définie par taxcov1(O,S)=DID/nombre de concepts de l'ontologie
- la couverture du corpus sur O définie par corpcov1(O,S)= TC/ le nombre de termes extraits du corpus

Les indicateurs suivants visent à évaluer la répartition des concepts de l'ontologie par rapport à la représentativité des concepts dans les documents. Aucun détail n'est donné sur l'affectation du poids d'un concept pour un document. Il est réalisé par le logiciel SemoiTaxonomy. W(c) correspond au poids général d'un concept dans le corpus S, $W_{di}(c)$ correspond au poids du concept c dans le document d_i et $nbc(d_i)$ correspond au nombre de concepts retrouvés dans le document d_i.

- La couverture de l'ontologie sur le corpus est calculée par $\text{taxcov}(O,S)=\frac{\sum_{di\in S, c\in O} Wdi(c)}{\sum_{c\in O} W(c)}$
- La couverture du corpus sur l'ontologie est calculée par $\text{corpcov}(O,S)=\frac{\sum_{di\in S, c\in O)} Wdi(c)}{\sum_{di\in S} S(TDi)}$

A partir de ces différents facteurs, une application a été développée pour comparer une ontologie par rapport à différents documents et à différents corpus. Elle permet de regrouper des documents ou corpus partageant les mêmes valeurs par ces mesures. Ces indicateurs permettent également de rechercher dans la hiérarchie de concepts des sous-parties de l'ontologie qui ont des valeurs pour ces indicateurs plus importantes que les parties qui les entourent.

Cette approche est originale et présente l'avantage d'avoir été évaluée dans plusieurs contextes (veille technologique et gestion de projet). Elle pourrait être utilisée à des fins de RI car elle permet de comparer différentes ontologies par rapport aux indicateurs, et pourrait servir à choisir l'ontologie. Cependant, des détails sur l'appariement des concepts de l'ontologie aux documents et la mise en place des mesures définies sont insuffisants pour évaluer totalement son

intérêt. De plus, elle repose sur des hiérarchies de concepts, ce qui, de notre point de vue, limite le potentiel sémantique de la représentation qui peut être utile pour la RI.

[Brewster 2004] propose également une approche visant à choisir entre plusieurs ontologies légères existantes pour une application donnée. Le principe de cette méthode consiste à confectionner un corpus de référence abordant le domaine de connaissance lié à l'application. Les ontologies sont ensuite comparées au corpus afin de déterminer l'ontologie la plus représentative de la connaissance recherchée. Une première comparaison entre le corpus et l'ontologie se fait par le calcul du taux de couverture du lexique de l'ontologie par rapport au lexique du corpus. Les termes du corpus sont extraits par un algorithme de regroupement reposant sur la sémantique latente (voir section 4.2) et sont étendus par l'intermédiaire d'hyperonymes issus de WordNet. Le lien entre les termes composant les regroupements et les concepts de l'ontologie est réalisé manuellement. Afin d'analyser la structure de l'ontologie, la mesure proposée par Stevenson [Stevenson 2002] est utilisée. Cette mesure a pour but d'évaluer si les concepts assemblés dans les mêmes regroupements sont aussi plus proches dans l'ontologie que ceux appartenant à des regroupements différents. Elle s'appuie sur une approche probabiliste prenant en compte l'occurrence des labels dans les groupes formés. Les labels considérés sont ceux des concepts de l'ontologie ainsi que ceux généralisant ces concepts dans WordNet.

Cette approche est originale et pourrait être réutilisée à des fins de RI car elle repose sur la mise en adéquation d'une ontologie avec un corpus. Cependant, seuls les résultats sur la comparaison de quatre ontologies et d'un corpus sont donnés. Aucune argumentation ne justifie les résultats obtenus ni les éventuelles méthodes qui permettraient de les optimiser. L'approche présente aussi plusieurs inconvénients. L'appariement en termes et concepts est réalisé manuellement. Le lexique extrait du corpus est augmenté par les termes hypernomyes dans WordNet. Cette ressource influence donc la qualité de l'évaluation. Finalement, seules les relations taxonomiques de l'ontologie sont considérées.

3.3 Bilan

La réutilisation d'une ontologie est un enjeu important compte tenu du coût de construction d'une telle ressource. Elle est difficilement envisageable dans le cas des ontologies lourdes très liées à la tâche pour laquelle elles sont construites. Dans le cas des ontologies légères, la réutilisation est une question qui a été étudiée dans la littérature. Différentes approches ont été mises en place pour évaluer de façon qualitative ou quantitative la réutilisabilité des ontologies légères. L'approche quantitative consiste à évaluer la réutilisabilité d'une ontologie par rapport à son efficacité dans la réalisation d'une tâche donnée. L'évaluation de la réutilisabilité d'ontologie pour la RI se place dans ce contexte. Les ontologies utiles pour la RI doivent être adaptées à la tâche de RI considérée et plus particulièrement apporter de la connaissance utile pour l'interprétation et la compréhension par le système des informations contenues dans le corpus documentaire.

L'ensemble des approches proposées dans la littérature pour évaluer l'adéquation d'une ontologie à un corpus considèrent uniquement les hiérarchies de concepts. Les relations présentes dans une ontologie peuvent être de plusieurs types (taxonomiques, méronymiques, causales, fonctionnelles ou sans propriété logique). Nous considérons que ne pas prendre en compte ces types de relation revient à ignorer une partie de la sémantique contenue dans l'ontologie. Nous proposons dans le chapitre 4 une méthodologie et des méthodes visant à pallier ces limites.

4 Indexation à partir d'ontologies

L'indexation de granules documentaires consiste à extraire les descripteurs représentatifs de leur contenu. Ces descripteurs sont ensuite utilisés pour permettre aux systèmes de rechercher les granules pouvant intéresser l'utilisateur.

L'indexation classique peut reposer sur différents types d'approches reposant soit sur une analyse lexicale, soit sur une analyse syntaxique du contenu des granules et prenant en compte uniquement les occurrences des mots dans le contenu. L'indexation à partir de la sémantique latente est une approche statistique qui vise à extraire la sémantique implicite contenue dans les documents. Cette approche regroupe les termes ayant des caractéristiques communes dans les documents et considère que les regroupements sont des unités de sens.

L'indexation sémantique repose sur une nouvelle intuition selon laquelle le sens des informations textuelles (et des mots qui composent les granules) dépend des relations conceptuelles entre les objets du monde auxquels elles font référence plutôt que des relations linguistiques et contextuelles trouvées dans leur contenu [Haav 2001]. L'indexation sémantique repose alors sur l'utilisation d'ontologies modélisant la conceptualisation des objets cités dans le corpus.

Nous détaillerons les différents procédés liés à l'indexation classique. Puis nous présenterons les travaux portant sur l'indexation sémantique.

4.1 Indexation automatique classique

L'indexation automatique consiste à déterminer automatiquement les termes représentatifs du contenu des granules documentaires dans l'objectif d'une recherche ultérieure. Les termes choisis doivent permettre de retrouver les granules documentaires pertinents par rapport à la représentation du besoin de l'utilisateur. Ces termes sont ensuite pondérés pour refléter leur pouvoir discriminant.

Le **langage de représentation** des granules contient l'ensemble des termes susceptibles d'être retenus pour représenter les granules. Ce langage peut être une sélection de termes choisis dans le contenu des granules ou bien il peut être constitué a priori ; on parle alors de langage contrôlé. Un vocabulaire contrôlé, un glossaire, une hiérarchie informelle ou un thésaurus (cf les ressources terminologiques décrites dans la section 4.2.2 du chapitre 1) peuvent être utilisés. Cependant, l'utilisation d'un thésaurus pour l'indexation n'a pas donné de bons résultats dans les expérimentations réalisées dans le cadre de TREC [SparkJones 1996].

Dans le cas où le langage de représentation est issu du corpus, plusieurs traitements peuvent être effectués sur le contenu des granules afin de sélectionner les termes de ce langage. Plus spécifiquement, l'approche classique repose sur la suppression des termes issus d'un anti-dictionnaire et la radicalisation des termes restants (nous avons décrit ce principe dans la section 2.4 du chapitre 2). Alternativement, les syntagmes extraits par des méthodes linguistiques peuvent être utilisés même si les résultats lors de la phase de recherche sont similaires aux résultats obtenus à partir de termes simples [Mitra 1997] [Hernandez 2003].

Le **pouvoir discriminant** des termes est ensuite calculé. Un terme apparaissant dans peu de documents est très pertinent pour décrire le contenu du granule car il permet de discriminer les documents entre eux. De plus, la fréquence d'un terme à l'intérieur du granule lui-même permet de révéler l'importance que l'auteur du granule a voulu donner à ce terme. Ce terme décrit a priori le contenu informationnel du granule. Ces remarques ont amené Robertson [Roberston 1976] à proposer une formule de pondération de l'importance d'un terme dans un document par rapport à l'ensemble des documents. Cette mesure est appelée tf.idf et a été reprise sous différentes versions dans la majorité de moteurs de recherche [Mothe 2000].

$$Tfidf_{i,j} = tf_{i,j} \times idf_j$$

$$idf_j = \log\left(\frac{N}{f_j}\right) + 1$$

où tf_{ij} est la fréquence du terme j dans le document i, N est le nombre total de documents , f_i est le nombre de documents contenant le terme j.

L'indexation repose donc généralement sur des mesures statistiques portant sur l'apparition des mots dans les granules documentaires. Une nouvelle génération de méthodes vise à prendre en compte les concepts abordés plutôt que les mots.

4.2 Indexation par la sémantique latente, vers une indexation conceptuelle

Les auteurs de la méthode statistique de la sémantique latente supposent que les textes sont porteurs d'une structure sémantique implicite dont ils tentent d'extraire les concepts en tant qu'unité de sens [Dumais 1990]. L'approche a pour but d'éviter la polysémie et la synonymie des termes retenus comme descripteurs par les approches statistiques classiques en regroupant les termes ayant des caractéristiques communes dans leur apparition dans les documents. La méthode a été créée à l'origine pour permettre une représentation des documents pouvant s'appliquer à des collections spécifiques en s'adaptant aux variations lexicales.

La méthode utilise comme données une matrice représentant les documents sur les colonnes et les termes sur les lignes. Pour la ligne i et la colonne j, la valeur représentée est la fréquence du terme i dans le texte j. La technique de décomposition en valeurs singulières permet de réduire cette matrice dans un espace réduit de dimensions orthogonales. L'originalité de la méthode est de réduire les dimensions de l'espace en modélisant les variations sémantiques significatives tout en diminuant le bruit.

Soit *Y*, la matrice représentant les termes de tous les documents.

Y peut s'écrire sous la forme d'une matrice rectangulaire comme le produit de deux matrices carrées et ayant l'avantage d'avoir les mêmes valeurs propres non nulles : $\mathbf{Y_{T,D} = Y_{T,T} * T_{D,D}}$

Y peut être décomposée en trois matrices :

- ***Doc***, matrice des documents,
- ***Term***, matrice des termes,
- ***Diag***, matrice diagonale des valeurs propres qui correspond aux concepts retrouvés dans les documents

La formule de reconstitution de Y est : $\mathbf{Y_{T,D} = Term_{T,T} * Diag_{T,D} * Doc^{t}_{M,D}}$

avec : *M* la dimension de la matrice, *T* le nombre de termes et D le nombre de documents.

Les valeurs propres de la matrice *Diag* sont rangées suivant leur pertinence décroissante. De cette matrice, les plus grandes valeurs propres sont gardées jusqu'au rang k pour former la matrice ***Diag'***.

Les matrices *Term* et *Doc* sont transformées. Seules les colonnes et les lignes correspondant aux valeurs propres de *Diag'* sont conservées pour former les matrices ***Term'*** et ***Doc'***.

Soit *Y'* telle que $\mathbf{Y'_{T,D}=Term'_{T,T} * Diag'_{T,D} * Doc'^{t}_{M,D}}$

Y' est l'unique matrice de rang k la plus proche de Y au sens des moindres carrés.

Les termes utilisés pour l'indexation peuvent être de simples mots ou des syntagmes nominaux déterminés par une procédure semi-automatique. Après la décomposition, les documents et les termes sont regroupés en collections en fonction de leur proximité spatiale.

L'indexation par sémantique latente a été évaluée dans la campagne d'évaluation TREC[1] et a donné de meilleurs résultats que les approches statistiques classiques telles que SMART [Dumais 1994] [Dumais 1995].

Le principal inconvénient de cette méthode est que la matrice Diag ou matrice des concepts n'est pas compréhensible par les humains. Ceci limite le pouvoir de la méthode qui ne décèle pas explicitement les concepts et donc la sémantique associée aux documents. Un autre inconvénient de la méthode est la complexité de l'algorithme qu'elle sous-entend. Elle est difficilement applicable à de larges collections car la taille des matrices augmente considérablement plus le nombre de documents et le nombre de termes sont élevés.

4.3 Indexation sémantique

Un autre type d'approche d'indexation vise à s'appuyer sur des ontologies pour représenter les granules documentaires ; ce type d'indexation s'appelle l'indexation sémantique. L'indexation sémantique repose sur l'intuition suivant laquelle le sens des informations textuelles (et des mots qui composent les granules) dépend des relations conceptuelles entre les objets du monde auxquels elles font référence plutôt que des relations linguistiques et contextuelles trouvées dans leur contenu [Haav 2001]. L'indexation sémantique n'est possible que par l'existence et l'utilisation de ressources décrivant explicitement l'information correspondant aux objets. Nous distinguons deux types de démarches dans l'indexation sémantique : la démarche issue de la RI et la démarche issue du Web Sémantique.

La démarche issue du domaine de la RI consiste à choisir comme langage de représentation des documents, l'ensemble des concepts et instances de l'ontologie. L'utilisation d'ontologies sous forme de hiérarchies de concepts, ontologies légères ou lourdes est le prolongement de l'utilisation dans le cadre de la RI des ressources terminologiques décrites dans le chapitre 1 [Haav 2001]. Les descripteurs ne sont plus choisis directement dans les documents ou dans un vocabulaire contrôlé (ou thésaurus) mais au sein même de l'ontologie. Les granules documentaires sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus [Aussenac 2004]. Il convient dans ce cas d'utiliser une ontologie reflétant le ou les domaines de connaissance abordés dans la collection documentaire. Il est en effet nécessaire de retrouver dans l'ontologie les concepts présents dans la collection pour indexer les documents à partir de toutes les thématiques abordées. Dans la littérature, il existe de nombreuses définitions de l'indexation sémantique. Certains auteurs différencient l'indexation sémantique de l'indexation conceptuelle [Mihalcea 2000]. L'indexation conceptuelle repose, pour eux, sur des hiérarchies de concepts ou ontologies de domaine, alors que l'indexation sémantique repose sur l'utilisation d'ontologies génériques telles que WordNet. L'ontologie WordNet étant, selon nous, limitée par rapport à la sémantique qu'elle peut contenir, nous ne considérons pas que les mécanismes d'indexation qu'elle permet de mettre en place soient plus « orientés sémantique ». Les ontologies de domaine peuvent par leur formalisation représenter des ressources impliquant

[1] http://trec.nist.gov/

un engagement sémantique plus fort (voir chapitre 1). Nous entendons donc par indexation sémantique, l'indexation de granules documentaires à partir de n'importe quelle ontologie de domaine. L'indexation sémantique se fait en deux étapes. La première étape consiste à identifier les concepts ou instances de l'ontologie dans les granules. La deuxième étape pondère les concepts pour chaque document en fonction de la structure conceptuelle dont ils sont issus [Haav 2001].

L'indexation sémantique est un type d'indexation qui s'inscrit également dans la démarche orientée Web Sémantique. Les précurseurs de cette nouvelle version du Web considèrent que les ressources participant au Web Sémantique seront toutes reliées entre elles par des relations sémantiques. Plus précisément, les données présentes sur le Web Sémantique seront modélisées sous forme d'ontologies où chaque ressource apparaît comme un élément de ces ontologies au même titre que la connaissance qui les décrit. L'objectif est donc d'ajouter au contenu du Web une structure formelle et de la sémantique (à travers des méta-données et de la connaissance) dans le but de permettre une meilleure gestion et un meilleur accès aux informations. Cette démarche repose sur des ontologies modélisant les objets du monde à travers les acteurs et entités que les documents constituent et comportent [Guha 2003]. Elles peuvent être vues comme une représentation des méta-données explicitement ou implicitement présentes dans les documents. La phase d'indexation est aussi appelée annotation de documents. L'annotation de documents a pour but de représenter les informations relatives au média (date de création, taille, format d'encodage), les méta-données présentes dans les documents (auteurs, date de production), les index (les descripteurs du contenu du document), l'identifiant du document par le système (emplacement) et une vue sur le contenu (résumé ou extraits) [Euzenat 2002]. La mise en place de cette nouvelle vision du Web dépend de la présence de ces méta-données. Un enjeu actuel du Web Sémantique est de définir des techniques permettant de les extraire [Kiryakov 2004] [Guha 2003]. La démarche orientée Web Sémantique a donc un double objectif : indexer le contenu des documents à partir des ressources permettant d'en extraire les concepts et instances mais aussi représenter les ressources en générant les méta-données correspondantes.

Ces deux approches sont pour nous complémentaires. Nous proposons donc de les combiner dans un modèle unique présenté dans le chapitre 4.

Les sections suivantes présentent les différents types d'ontologies pouvant servir à l'indexation sémantique des documents et les différentes étapes de l'indexation sémantique.

4.3.1 Différentes ontologies comme espace de représentation des documents

Différents types d'ontologies sont utilisés dans le cadre de l'indexation sémantique. Ces ontologies ne séparent pas les aspects de la connaissance liés au contenu des documents et ceux liés à la tâche de recherche réalisée.

WordNet est une ontologie souvent utilisée [Khan 2002] [Gonzalo 1998] [Mihalcea 2000] [Cucchiarelli, 2004]. La raison principale qui motive son utilisation est qu'elle a pour objectif de représenter la langue naturelle. Les approches présentées dans [Gonzalo 1998] [Mihalcea 2000] ont pour but d'identifier les descripteurs des documents dans l'ensemble des synsets de WordNet (termes synonymes définissant un sens d'un mot). Les approches présentées dans [Khan 2002] et [Cucchiarelli 2004] visent quant à elles à déterminer des parties aussi appelées « regions » de l'ontologie permettant de représenter un ensemble de documents donnés. Ces approches s'apparentent à la méthode présentée dans [Baziz 2005] et consistent à créer un réseau sémantique pour chaque document. L'inconvénient majeur de l'utilisation de WordNet est que cette ontologie est trop générale et peu formalisée pour modéliser correctement un domaine donné.

D'autres méthodes, au contraire, s'appuient sur des ontologies de domaine, ce qui permet de mieux spécifier le langage d'indexation. La hiérarchie de concept MESH est utilisée pour

indexer des documents de la médecine dans [Hearst 1997]. Le projet Menelas vise à développer un système permettant d'accéder aux rapports médicaux de centres hospitaliers. Il repose donc sur une ontologie construite à partir des rapports à indexer qui modélise l'ensemble des maladies coronariennes [Zweigenbaum 1993].

Les approches orientées Web Sémantique s'appuient quant à elles sur des ontologies visant à représenter l'ensemble des méta-données qui peuvent être associées aux documents. Ces ontologies sont formelles et permettent de mettre en place des inférences à partir de leurs axiomes. Dans (Ka) [2] [Benjamins 1999] des pages Web concernant des chercheurs du domaine de l'acquisition des connaissances sont manuellement annotées à partir des concepts d'une ontologie. L'ontologie contient des éléments décrivant les personnes, les organisations, les publications ainsi que le domaine de l'acquisition de connaissance et les domaines scientifiques connexes.

Le modèle proposé dans [Vallet 2005] repose sur le même principe mais organise les concepts de l'ontologie à partir de quatre classes de haut niveau. Certaines classes représentent les éléments de contenu des documents, d'autres représentent les méta-données. Cependant, la connaissance n'est pas distinguée par rapport à son utilité dans la tâche. Ainsi, la réutilisabilité de l'ontologie pour une autre tâche ou un autre domaine n'est pas assurée. De la même façon, les éléments relatifs au domaine thématique du corpus et ceux relatifs à la tâche ne sont pas distingués dans le système présenté dans [Guha 2003]. Ce système indexe un ensemble de documents extraits du Web couvrant plusieurs domaines tels que la musique, la météo, les sites d'achat en ligne. Il repose sur une ontologie définissant les principales entités du domaine telles que les personnes, les endroits, les événements, les organisations et les documents. Le même problème se retrouve dans l'approche présentée dans [Kiryakov 2004].

Les sections suivantes indiquent comment une ontologie est utilisée pour l'indexation de documents. Elle repose sur deux étapes : l'identification des concepts et instances dans les documents et leur pondération.

4.3.2 Identification des concepts et des instances existant dans l'ontologie

La première étape de l'indexation conceptuelle consiste à identifier les concepts et/ou instances de l'ontologie apparaissant dans les granules.

Une première approche consiste à identifier ces éléments de l'ontologie manuellement dans les documents. Cette approche suivie dans [Vallet 2005] [Paralic 2003] [Kahan 2001] est généralement réalisée par un expert et a pour intérêt d'être fiable car l'expert interprète la sémantique associée aux concepts dans l'ontologie et choisit le concept représentant au mieux la notion abordée dans le document. Cependant, même assisté par des traitements automatiques, ce procédé reste fastidieux, coûteux en temps et implique des erreurs [Erdmann 2000].

D'autres approches visent à automatiser ce procédé. Cette démarche est légitime dans la mesure où l'utilisation d'une ontologie permet d'accéder à la connaissance et de la rendre manipulable par les systèmes. Dans ce cas là, les labels ou termes désignant les concepts ou instances sont recherchés dans les granules documentaires. Un concept (et une instance de concept) est en effet défini à partir d'un ou plusieurs labels représentant les variantes lexicales que peuvent prendre les termes définissant les concepts [Vallet 2005] [Kiryakov, 2004] [Guha 2003].

4.3.2.1 Extraction des termes du granule

L'approche généralement suivie consiste à extraire des documents l'ensemble des termes y apparaissant et d'y rechercher les labels contenus dans l'ontologie. L'extraction de termes des documents se fait de la même façon que la recherche du langage de représentation classique. Les termes apparaissant dans un anti-dictionnaire peuvent être supprimés. Les expressions sont

extraites soit statistiquement, soit syntaxiquement. L'extraction d'expressions est quasiment obligatoire car les labels des concepts sont souvent composés de ce type d'éléments.

4.3.2.2 Recherche des labels correspondant à des concepts ou instances de l'ontologie

Les labels sont recherchés dans l'ensemble des termes extraits en favorisant la prise en compte des labels les plus longs et donc des concepts les plus spécifiques [Bloehdorn 2004] [Baziz 2005] [Vallet 2005]. Par exemple, dans le cas où les labels « Madrid », « Real », et « Real Madrid » apparaissent dans le document, le label retenu - et donc le concept correspondant, - sera Real Madrid car l'expression formée de deux termes est plus précise que le ou les termes seuls. Plusieurs algorithmes ont été définis pour rechercher les labels les plus longs, ils consistent à faire varier la taille d'une fenêtre sur les mots de chacune des phrases des textes.

4.3.2.3 Désambiguïsation des labels

Les labels peuvent cependant se rapporter à plusieurs concepts. Dans ce cas, un mécanisme de désambiguïsation du terme est mis en place afin d'identifier quel est le concept abordé dans le document. Il existe un grand nombre de techniques de désambiguïsation [Sanderson 2000]. Les premières études faites sur l'intérêt d'utiliser la désambiguïsation en RI ont amené à des résultats variés, voire même contradictoires. Cependant, la conclusion qui peut être tirée de ces expériences est que des algorithmes de désambiguïsation de haute qualité sont nécessaires pour améliorer les performances du système [Sanderson 2000]. Les techniques les plus simples considèrent les approches suivantes. La stratégie du « tout » correspond au cas dans lequel tous les concepts sont considérés. La stratégie du « premier » consiste à restituer le concept le plus fréquent dans le document ou bien dans la collection. La stratégie du « contexte » base la désambiguïsation sur la proximité sémantique des concepts candidats et du contexte dans lequel ils apparaissent dans les documents. Cette dernière variante peut être mise en place de diverses façons. Des règles syntaxiques et lexicales peuvent être générées manuellement, elles déterminent le sens d'un mot à partir des termes qui lui succèdent ou le précèdent dans son contexte [Small 1982]. Cette approche a l'inconvénient de ne permettre la désambiguïsation que d'une faible proportion de termes. La désambiguïsation peut aussi s'appuyer sur des corpus déjà désambiguïsés. C'est le cas d'une des stratégies suivies dans [Mihalcea 2000]. Le contexte du terme est représenté par les expressions qu'il forme à partir de tous les termes qui apparaissent directement après lui dans le corpus et directement avant lui. Le sens du mot est alors choisi à partir de son sens le plus courant dans les expressions représentant son contexte dans le corpus de référence SemCor [Miller 1993]. La limite de cette approche est que peu de ressources existent et qu'elles ne couvrent pas des domaines spécifiques. La désambiguïsation peut également reposer sur l'utilisation de ressources telles que des dictionnaires, des thésaurus ou des ontologies. L'utilisation de dictionnaires a pour principe de comparer les termes formant les différentes définitions du terme à désambiguïser avec les termes apparaissant dans le contexte du terme polysémique. Cette approche est suivie notamment dans [Lesk 1988] et [Mihalcea 2000]. Dans [Mihalcea 2000], le contexte d'un mot est représenté par les mots qui l'encadrent dans les documents dans une fenêtre de dix mots. Le sens choisi est celui dont la définition contient le plus de mots du contexte. Les relations entre termes (synonymies, est lié à) présentes dans les thésaurus et WordNet sont aussi utilisées pour désambiguïser les termes. Dans le cas où les ressources sont organisées hiérarchiquement, les mesures de similarités entre concepts telles que celles présentées dans la section 1 peuvent être utilisées [Banerjee 2002] [Patwardhan 2003].

4.3.2.4 Extraction de nouvelles instances

L'extraction d'instances de concepts a pour but d'extraire les méta-données qui permettront de représenter les ressources dans le cadre du Web Sémantique. L'extraction d'instances repose sur des techniques du domaine de l'extraction d'information. De nombreuses

plate-formes telles que Gate [Cunningham 2002] permettent de définir des patrons d'extraction ou d'utiliser des techniques reposant sur le traitement automatique des langues.

L'extraction d'instances de concepts peut se faire à partir de techniques d'extraction d'entités nommées, issues du domaine du traitement automatique des langues [Kiryakov 2004]. Une entité nommée est un nom ou syntagme nominal se rapportant à une entité comme, par exemple, une personne, une organisation ou une localisation [Chinchor 1998]. Un procédé d'extraction d'instances est décrit dans [Kiryakov 2004]. Les entités sont extraites à partir d'une base de connaissance qui, à partir de ressources lexicales, permet la détection automatique des entités. Les ressources lexicales décrivent par exemple les suffixes pouvant permettre la détection de noms d'entreprises ou de noms de familles ou de personnes. La base de connaissances contient un ensemble d'instances prédéfinies et décrites à partir d'axiomes. Un mécanisme d'inférence définit des règles permettant d'extraire de nouvelles instances. L'utilisation d'entités nommées et d'instances d'ontologie est une approche originale car les entités nommées sont rarement considérées en RI. La raison qui motive ces travaux est qu'une étude récente, faite sur les SRI, montre que 25% des requêtes contiennent des noms de personnes [Dumais 2003]. De plus, dans une approche traditionnelle par mot clé, l'utilisateur est obligé de spécifier à la fois le mot désignant l'instance qu'il recherche ainsi que les concepts auxquels se rapporte l'instance afin d'affiner sa recherche. Dans une approche conceptuelle, cette information n'a pas besoin d'être précisée car elle est connue par le système.

L'étape suivante consiste à pondérer les termes afin de mesurer leur représentativité du document.

4.3.3 Pondération des concepts et instances

Le calcul du poids d'un concept ou d'une instance dans la représentation d'un granule peut être fait suivant plusieurs approches : statistiques ou conceptuelles.

4.3.3.1 Pondération statistique

L'approche proposée dans [Vallet 2005] a pour but de calculer le poids des instances. Elle est inspirée de la méthode tf.idf.

Le poids $w_{i,j}$ d'une instance I_i dans un document D_j est calculé ainsi :

$$wi,j = \frac{freqi,j}{\max k\ freqk,j} * \log \frac{N}{ni}$$

où $freq_{i,j}$ représente le nombre d'occurrences de I_i dans D_j, $\max_k freq_{k,j}$ est la fréquence de l'instance dans D_j, n_i est le nombre de documents annotés avec I_i et N est le nombre total de documents dans la collection.

Le nombre d'occurrences d'une instance a été défini comme le nombre de fois où le label de l'instance apparaît dans le texte, si ce document est annoté avec l'instance, ou 0 s'il ne l'est pas. Cependant, les résultats obtenus n'ont pas été satisfaisants car un grand nombre d'instances n'était pas reconnu à cause de lacunes à l'étape précédente correspondant à l'extraction des labels (non prise en compte des pronoms et périphrases notamment). Une approche similaire est présentée pour la pondération de concepts dans [Baziz 2005]. L'inconvénient de ces approches est qu'elles ne considèrent que les occurrences des concepts ou instances dans les documents et ne considèrent pas l'organisation conceptuelle dont ils sont issus. Une partie de la sémantique contenue dans les relations entre concepts est alors ignorée.

D'autres approches visent à combiner la pondération des concepts et/ou instances à partir de leurs occurrences dans les documents et leur place dans la représentation conceptuelle. Elles reposent sur le calcul de similarité entre concepts présenté dans la section 1.

4.3.3.2 Pondération à partir de similarité conceptuelle

Dans [Desmontils 2002] une approche est présentée pour indexer un ensemble de sites Web à partir d'une ontologie. Le pouvoir représentatif d'un concept prend en compte la fréquence d'apparition des termes désignant le concept dans les sites mais également ses relations avec les autres concepts du domaine. Plus un concept a des relations avec les autres concepts présents dans la page, plus il est représentatif de la page. Le pouvoir se calcule de la façon suivante :

Les termes d'une page Web sont tout d'abord extraits après analyse syntaxique (tree tagger) à partir de patrons (nom, nom+nom, nom+adjectif). Un premier poids, appelé poids de fréquence est calculé pour chaque terme en fonction de sa fréquence d'apparition et des balises html qui l'encadrent. Les coefficients correspondant à chaque balise sont attribués expérimentalement, par exemple, si un terme est encadré par la balise titre, le coefficient est 10, s'il est mis en gras, le coefficient est 2.

En supposant qu'un terme T_i apparaît p fois dans une page contenant n termes, $M_{i,j}$ étant le coefficient relatif à la balise encadrant l'occurrence j du terme T_i, le poids de fréquence P_freq de T_i est calculé ainsi :

$$P\_freq(T_i)=\frac{P(T_i)}{\max_{k=1..n}(P(T_k))} \quad \text{et } P(T_i) = \sum_{i=1}^{p}(Mi,j)$$

Ensuite, à partir de WordNet, l'ensemble des concepts relatifs à ces termes est généré sous forme de synset en prenant tous les sens définis. Un poids, appelé poids sémantique, est ensuite calculé en mesurant la similarité entre le concept donné et l'ensemble des autres concepts retrouvés.

Pour calculer la similarité entre 2 concepts, la formule sim définie dans [Wu & Palmer 94] est utilisée :

$$\text{Sim}(c1,c2)=\frac{2*depth(c)}{depth(c1)+depth(c2)}$$

où depth(cc) correspond au niveau de profondeur du concept cc dans la hiérarchie

et c est le concept subsumant c1 et c2.

Pour plus de détails sur la mesure de similarité voir la section 1.

Pour calculer le poids sémantique d'un concept dans une page, la somme des mesures de similarité du concept avec les autres concepts retrouvés de la page est calculée de la façon suivante :

$$P\_sem(synset_i(T_k))=\sum_{j\in[1,k-1]\cup[k+1,m]}\sum_{l=1}^{k} sim(synset_i(T_k),synset_l(T_j))$$

où $synset_i(T_m)$ représente le sens i dans WordNet retrouvé pour le terme T_m

Enfin, le pouvoir représentatif Rep du concept ou synset correspondant aux termes T_k est calculé en fonction de son poids sémantique et de son poids de fréquence :

$$\text{Rep}(\text{synset}(T_k))=\frac{\alpha*P\_freq(Tk)+\beta*P\_sem(symset(Tk))}{\alpha+\beta}$$

α et β sont fixés empiriquement à 1 et 2.

Les concepts retenus pour indexer chaque page sont ensuite choisis à partir d'un seuil sur ce pouvoir et de la présence de ce concept dans l'ontologie choisie pour indexer le corpus.

4.4 Bilan

Dans le contexte de la RI, une ontologie représente la connaissance utile pour permettre une meilleure indexation. Il est donc indispensable qu'elle possède une forte composante lexicale afin de pouvoir mettre en correspondance les contenus des documents et les labels des concepts. Les mécanismes de pondération sont repris des mécanismes classiques de RI mais l'indexation est réalisée au niveau des concepts et non plus au niveau des termes.

Dans le contexte du Web Sémantique, les ontologies sont décrites formellement. Le principe suivi consiste à reposer sur une ontologie dans laquelle l'ensemble des concepts est défini et la phase d'indexation consiste à extraire des instances des concepts dans les documents.

5 Accès aux documents à partir d'ontologie

La plupart des SRI fonctionnent avec une interface qui permet à l'utilisateur de formuler son besoin en informations à partir d'une requête. Le système lui présente le résultat de sa recherche sous forme d'une liste de références vers les documents retrouvés. Une alternative au principe de recherche d'information ad-hoc consiste à fournir des outils qui permettent à l'utilisateur d'explorer la collection de documents pour trouver les documents pertinents sans avoir à exprimer son besoin en informations sous forme d'une requête conventionnelle. L'utilisateur a en effet du mal à spécifier son besoin et à l'exprimer, surtout s'il ne connaît pas le contenu de la collection à sa disposition. Notre domaine d'étude se place dans ce contexte là.

Dans la section 4.1 nous décrivons les langages d'interrogation des SRI, les mécanismes de formulation de requête et d'appariement requête/granule. Ces points sont présentés dans les approches classiques, puis en utilisant des ontologies. Dans la section 4.2, nous présentons les mécanismes liés à l'exploration de corpus à partir de hiérarchies de concepts. La section 4.3 est consacrée à l'exploration de corpus à partir d'ontologie et la section 4.4 présente les mécanismes de navigation à base d'ontologies. Dans la dernière section, nous dressons un bilan de ces différentes approches.

5.1 Langage d'interrogation, requête et appariement

Afin de communiquer son besoin au système, l'utilisateur doit le formuler dans un langage interprétable par le système.

5.1.1 Interrogation en langage libre

Dans le cadre de la recherche d'information ad-hoc, ce besoin est formulé sous forme de requêtes. La formulation de la requête est un problème crucial car de sa qualité dépend la qualité des documents restitués par le SRI. Le format de la requête dépend du SRI. Les requêtes booléennes sont composées de termes et d'opérateurs booléens (ET, OU, SAUF). Les documentalistes maîtrisent mieux ce type de requête qui est souvent difficile à formuler pour un utilisateur non initié [Mothe 2000]. Ce type de requête est le plus utilisé pour l'accès à des bases spécialisées (Pascal, Questel). Il est également disponible pour de nombreux moteurs de recherche sur le Web tels que google[2] et yahoo[3] à partir d'interfaces de recherches avancées.

Un autre type de requête consiste à formuler les requêtes en langage libre. Aucune syntaxe particulière n'est alors définie. L'appellation langage libre est préférée à langage naturel car

[2] http://www.google.fr/advanced_search?hl=fr

[3] http://fr.search.yahoo.com/Web/advanced?fr=fp-top

généralement, les requêtes formulées par l'utilisateur ne constituent pas des phrases grammaticales correctes mais des listes de mots ou d'expressions. La majorité des moteurs de recherche proposent ce langage par défaut (google, yahoo, voila).

Plusieurs modèles visent à réaliser l'appariement entre la représentation des granules documentaires et la représentation des requêtes. Nous citons ici les modèles de base, pour plus de détail voir [Mothe 2000]. Le modèle booléen est basé sur l'algèbre de Boole et repose sur une représentation booléenne des requêtes. Dans ce modèle, les documents restitués à l'utilisateur sont ceux contenant exactement les termes de la requête. Il repose donc sur l'absence ou la présence des termes retenus pour indexer les documents et les termes de la requête. Le modèle vectoriel présenté par Salton repose sur les fondements mathématiques des espaces vectoriels [Salton 1971]. Dans ce modèle, les documents et les requêtes sont représentés sous forme de vecteurs dans l'espace des termes, issus de l'indexation. Les documents sont ensuite ordonnés à partir de leur ressemblance à la requête. Plusieurs mesures (Produit scalaire, Mesure de Dice, Mesure de Jaccard, …) permettent de calculer la similarité entre ces deux éléments correspondant aux calculs de la distance entre les deux vecteurs. Le modèle probabiliste [Roberston 1976] repose sur la probabilité de pertinence d'un document connaissant la requête. Le modèle de langage [Ponte 1998] mesure la probabilité de générer une requête à partir du modèle de langage du document. Dans l'ensemble de ces modèles, les documents sont restitués à l'utilisateur par ordre de pertinence supposée décroissante.

L'utilisation d'ontologies dans les SRI permet de définir d'autres types d'interrogation qui s'appuient sur les langages du Web Sémantique.

5.1.2 Interrogation à partir d'un langage dédié aux ontologies

Dans le cas où les documents sont représentés à partir d'ontologies lourdes, des moteurs d'inférences peuvent être intégrés au système afin d'interroger la base de connaissance constituée des ontologies et des documents. Racer et FaCT DL sont des exemples de moteurs d'inférence reposant sur la logique de description. Afin d'interroger ces moteurs, plusieurs langages d'interrogation ont été définis à partir du langage formalisant la connaissance [Karvounarakis, 2002] [McBride. 2001] [Miller 2002] [Guha 2003]. Ces langages fournissent des mécanismes permettant d'exprimer des requêtes complexes. La requête est alors exécutée sur la connaissance représentée dans l'ontologie et les instances qui satisfont la requête sont retournées. Les requêtes sont soit générées à partir d'une requête en langage libre [Guha 2003] [Rocha 2004], soit à partir d'interface permettant de sélectionner les classes et propriétés de l'ontologie qui intéressent l'utilisateur [Kiryakov 2004] [Maedche 2003]. L'avantage de ce type d'interrogation est que des mécanismes d'inférence sont mis en place à partir de la classification hiérarchique des concepts et des règles. GetData présenté dans [Guha 2003] permet d'interroger de façon simple et efficace une ontologie considérée comme un graphe étiqueté représenté en RDF. Il permet d'accéder à des ressources vérifiant une propriété.

L'avantage de ce type d'interrogation est qu'elle permet de mettre en place des procédés de raisonnement à partir des éléments de la requête et des éléments retrouvés dans les documents. Cependant, ce type d'appariement consiste à rechercher exactement les éléments présents ou inférés de la requête dans les documents. Les documents ne sont pas restitués par ordre de pertinence mais parce qu'ils contiennent les éléments cibles.

L'originalité de l'approche présentée dans [Vallet 2005] est que l'ensemble des documents correspondant aux instances retournées par une requête formulée en RDQL est ensuite classé par calcul de pertinence vis à vis de la requête. L'appariement repose sur la représentation vectorielle de la requête et des documents à partir des instances de l'ontologie, reprenant ainsi les principes développés par Salton.

5.1.3 Appariement à partir d'ontologies

Les ontologies peuvent servir à calculer la similarité entre la représentation de la requête et la représentation des documents dans le cas où les deux représentations sont faites à partir des concepts d'une même ontologie.

Cette approche est suivie dans [Andreasen 2003]. Les documents et requêtes sont représentés à partir du langage et de l'ontologie Ontologu. Cette ontologie contient un ensemble de concepts et de relations entre concepts, dont la relation de subsomption. Elle est considérée comme un graphe orienté. L'avantage du calcul de la similarité est de classer les documents restitués par rapport à leur similarité à la requête, cette similarité reposant sur l'organisation des concepts dans l'ontologie. Le calcul de similarité s'appuie sur trois intuitions.

- La première intuition est que les documents liés au concept généralisant ou spécifiant le concept utilisé dans le requête peuvent intéresser l'utilisateur. Le calcul de la similarité prend donc en compte la distance séparant les deux concepts par la relation de subsomption. La similarité revient à prendre le nombre d'arcs séparant les deux concepts par le chemin le plus court à partir de la relation de subsomption.
- La deuxième intuition est que deux concepts ayant un concept les généralisant (ou subsumeur) commun sont plus similaires. Afin d'appliquer cette intuition, chaque concept est représenté par un ensemble flou à partir des concepts le généralisant. La similarité entre concepts est alors calculée à partir des éléments faisant partie de l'intersection entre les descriptions des concepts.
- La troisième intuition est que la similarité entre concepts doit prendre en compte les relations autres que les relations de subsomption. L'ensemble des concepts généralisant les deux concepts est alors considéré. Un sous-graphe de l'ontologie est construit à partir des concepts de cet ensemble pouvant être reliés dans l'ontologie par n'importe quel type de relation. La similarité est calculée par rapport aux nombres de nœuds ainsi connectés.

Cette approche est originale car elle calcule la similarité entre les concepts des documents et les concepts de la requête. La mesure de similarité proposée repose sur l'organisation des concepts dans l'ontologie. Cependant, aucune indication n'est donnée sur la combinaison des différents facteurs de la mesure. De plus, les auteurs ne considèrent pas le cas de figure suivant lequel plusieurs concepts sont retrouvés à la fois dans la requête et les documents et comment les différentes similarités sont combinées. Aucune évaluation n'est proposée.

Une autre approche est présentée dans [Guarino 1999]. Contrairement à la précédente, le mécanisme d'appariement est entièrement décrit ; cependant le procédé est manuel. Le but du système OntoSeek [Guarino 1999] est d'améliorer l'accès aux pages jaunes à partir d'un mécanisme reposant sur WordNet. Les documents et les requêtes sont représentés à partir de graphes conceptuels formés de nœuds et d'arcs dont les labels sont issus de WordNet. Une interface aide l'utilisateur dans la conception de ces graphes. Pour étiqueter les nœuds, l'utilisateur peut soit proposer des mots qui sont ensuite désambiguïsés à partir de WordNet, soit directement naviguer dans WordNet pour sélectionner les synsets qui l'intéressent. De la même façon, les arcs sont étiquetés soit à partir d'une liste proposée par le système, soit à partir de termes proposés par l'utilisateur. Un procédé d'appariement entre le graphe de la requête et l'ensemble des graphes représentant les documents est ensuite mis en place. Le système recherche les graphes de documents qui subsument (ou qui spécifient) le graphe de la requête. Les résultats sont présentés à l'utilisateur à partir d'une interface présentant un rapport en HTML.

5.1.4 Reformulation de requête à partir des termes de l'ontologie

Il a été prouvé que la reformulation de requêtes a des effets positifs en RI [Harman 1992]. L'objectif de la reformulation est soit de limiter le silence (le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système) soit de réduire les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système). Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête initiale. Dans le second cas la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin. Il y a principalement deux approches permettant l'expansion de requêtes. La première consiste à utiliser des ressources, comme par exemple un dictionnaire [Moldovan 1999] ou bien WordNet [Voorhes 1994], en étendant les requêtes à partir de nouveaux termes en relation avec les termes de la requête. La deuxième solution est la ré-injection de pertinence reposant sur l'analyse des termes contenus dans les documents jugés pertinents pour la requête initiale. Cette approche, ne faisant pas intervenir d'ontologies, ne fait pas partie de notre étude.

Un autre intérêt des ontologies est de permettre la désambiguïsation des termes de la requête. Dans [Guha 2003] la désambiguïsation se fait selon trois approches. La première consiste à choisir le concept dont les labels apparaissent le plus dans les documents. La seconde approche consiste à réaliser un profil utilisateur et à choisir le concept le plus proche de son profil. Finalement, la troisième prend en compte le contexte de la recherche et les documents recherchés par l'utilisateur jusque là. Aucune étude comparative n'est présentée.

Dans $(Ka)^2$ [Benjamins 1999], les pages Web sont annotées manuellement par des concepts d'une ontologie. Pour une requête donnée, tous les concepts liés aux termes de la requête sont inférés et ajoutés à la requête. Une interface a été développée pour assister l'utilisateur dans la formulation ou le raffinement de sa requête. Elle repose sur la visualisation de l'ontologie à partir de vues hyperboliques. Il est ainsi possible de naviguer dans l'ontologie et de centrer la visualisation sur la représentation des concepts intéressant l'utilisateur comme il a été fait dans WebBrain[4].

5.2 Exploration à partir de hiérarchie de concepts

La catégorisation suivant une hiérarchie de concepts est une façon de décrire intuitivement l'information [Lawrie 2000]. Plusieurs interfaces d'exploration de collections documentaires reposent sur cette structuration. Elles visent à aider l'utilisateur dans la spécification de son besoin en lui donnant une vue d'ensemble sur la collection, puis en lui permettant de spécifier les vues en fonction des informations qui l'intéressent. Plusieurs types d'interfaces reposant sur des hiérarchies de concepts ont été développées. La particularité de chacune d'entre elles repose sur l'utilisation des hiérarchies pour proposer les vues sur la collection.

Cat-a-cone [Hearst 1997] est un système reposant sur une interface qui permet à l'utilisateur de consulter une collection de documents via la navigation d'une grande hiérarchie de catégories (MeSH) auxquelles les documents sont associés. L'utilisateur peut spécifier les concepts qu'il veut visualiser, ce qui induit une modification de l'organisation et de l'affichage de l'ensemble des documents. Une copie d'écran de l'interface est présentée dans la figure 3.4.

[4] http://www.Webbrain.com/html/default_win.html

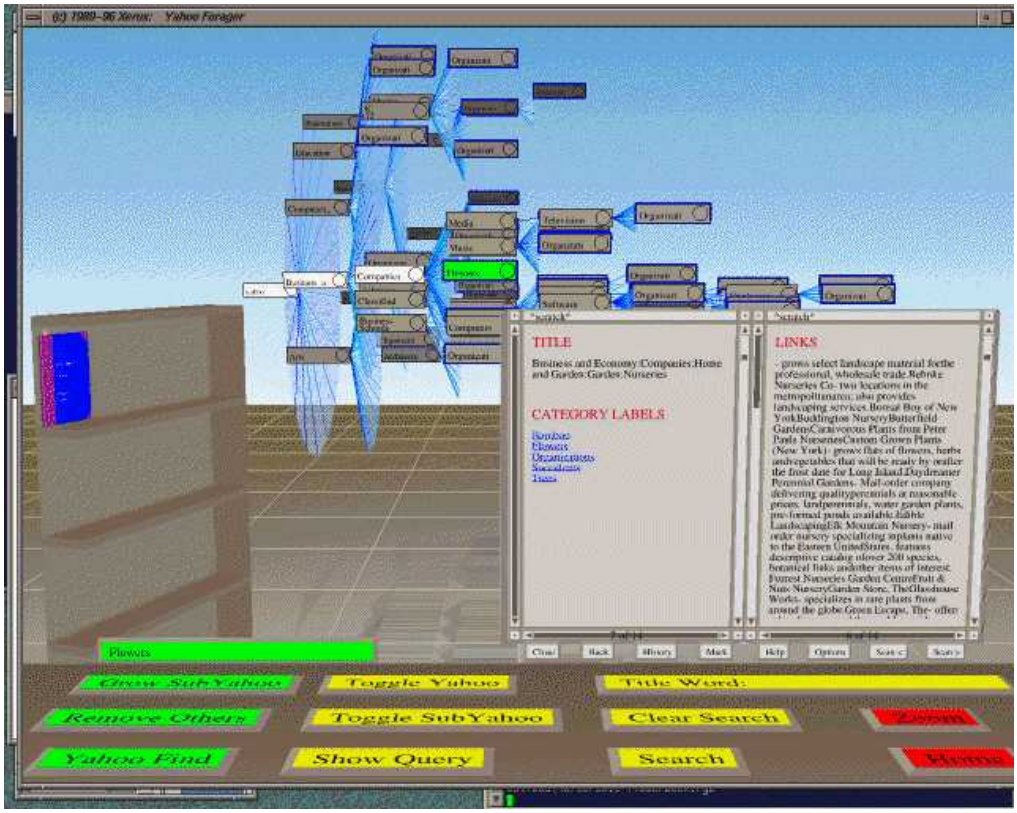
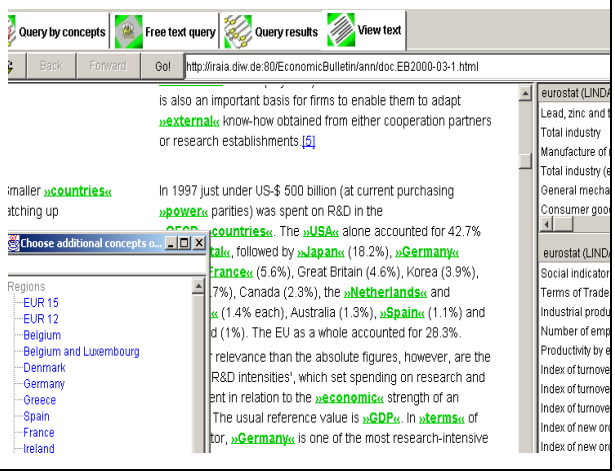


Figure 3.4 Interface du système Cat-a-Cone

Un autre type d'interfaces, telle que celle réalisée dans le cadre du projet IRAIA présentée dans [Mothe 2003b], permet à l'utilisateur de formuler et reformuler des requêtes à partir de l'exploration des différentes hiérarchies de concepts selon lesquelles les documents sont indexés. Les documents correspondant aux concepts sélectionnés dans la requête sont ensuite présentés par regroupements à partir de l'ensemble des concepts qu'ils partagent. Différentes copies d'écran du projet IRAIA sont données dans la figure 3.5.

Interrogation via des hiérarchies de concepts

Reformulation de requête

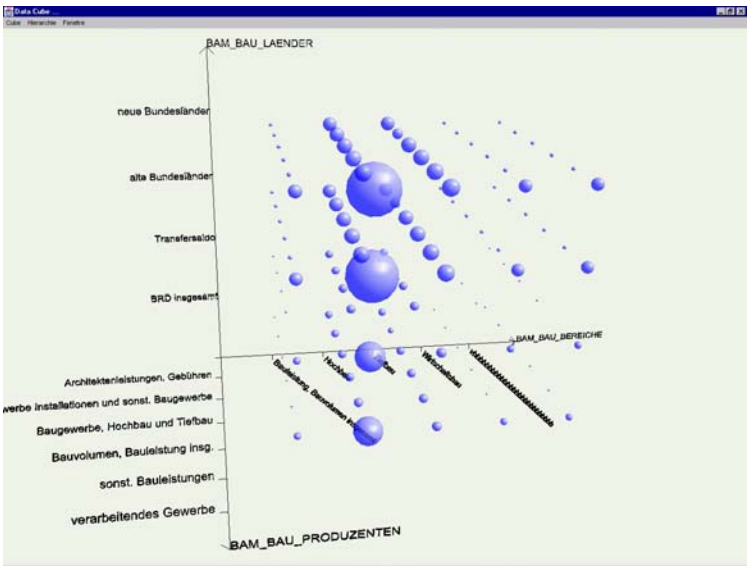


Visualisation des résultats par groupes

Figure 3.5 Interface d'IRAIA

Un autre type d'interface repose sur une analyse multi-dimensionnelle des documents en considérant que les hiérarchies de concepts utilisées par le système sont des dimensions de la collection. Le système DoCube [Mothe 2003a], dont une capture d'écran est présentée dans la figure 3.6, permet ce type d'analyse. L'utilisateur peut analyser le nombre de documents partageant plusieurs concepts décrivant différents aspects de la collection. Il peut ainsi choisir d'utiliser les concepts dans une requête en ayant une idée du nombre de documents qui y correspondent dans la collection

Vue 3D (selon 3 hiérarchies de concepts)



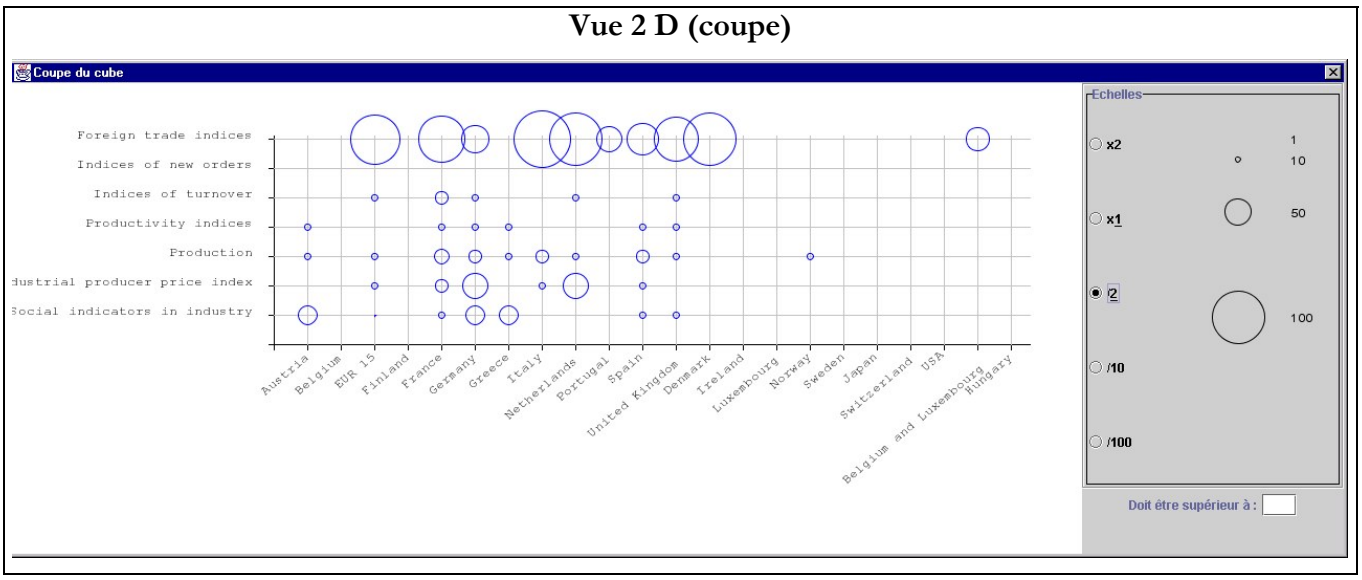


Figure 3.6 Interface DocCube

5.3 Exploration à partir d'ontologies

Les approches présentées dans [Seeling 2003] [Stuckenschmidt 2004] visent à permettre l'exploration de corpus documentaires en regroupant les documents partageant des méta-données à partir de leur représentation sous forme d'ontologies.

Les travaux présentés dans [Stuckenschmidt 2004] consistent à indexer automatiquement un ensemble de publications à partir de mots clés trouvés dans un thésaurus de sciences naturelles (ressource terminologique composée de relations hiérarchiques entre les termes, de relations de préférences, et des relations « est une drogue liée à » « est une maladie liée à »). Les index ainsi détectés mais aussi le titre, le ou les auteurs et les informations relatives à la publication de l'article sont ensuite stockés dans une ontologie en tant que méta-données du document. Pour chaque document, un concept est créé dans l'ontologie, ce concept ayant un ensemble de propriétés (nom des auteurs, titre…) et un ensemble de liens vers des termes du thésaurus (termes retenus pour l'indexation). L'exploration de la collection repose ensuite sur le regroupement de documents partageant les mêmes méta-données. Une interface permet de visualiser ces regroupements en présentant l'ensemble des groupes de documents et le terme du thésaurus de l'ontologie ayant permis ces regroupements. Elle est présentée dans la figure 3.7. La partie en haut à gauche de l'interface permet de centrer l'exploration sur un terme issu du thésaurus. Au-dessous, l'ensemble des termes co-occurrants dans les documents avec le terme considéré est présenté. L'utilisateur peut ainsi se focaliser sur un nouveau terme. Dans la partie droite, les regroupements des documents partageant le terme sont présentés. L'utilisateur peut ainsi évaluer les méta-données communes aux documents traitant du terme. Il peut aussi accéder au contenu du document en cliquant sur une instance d'un groupe.

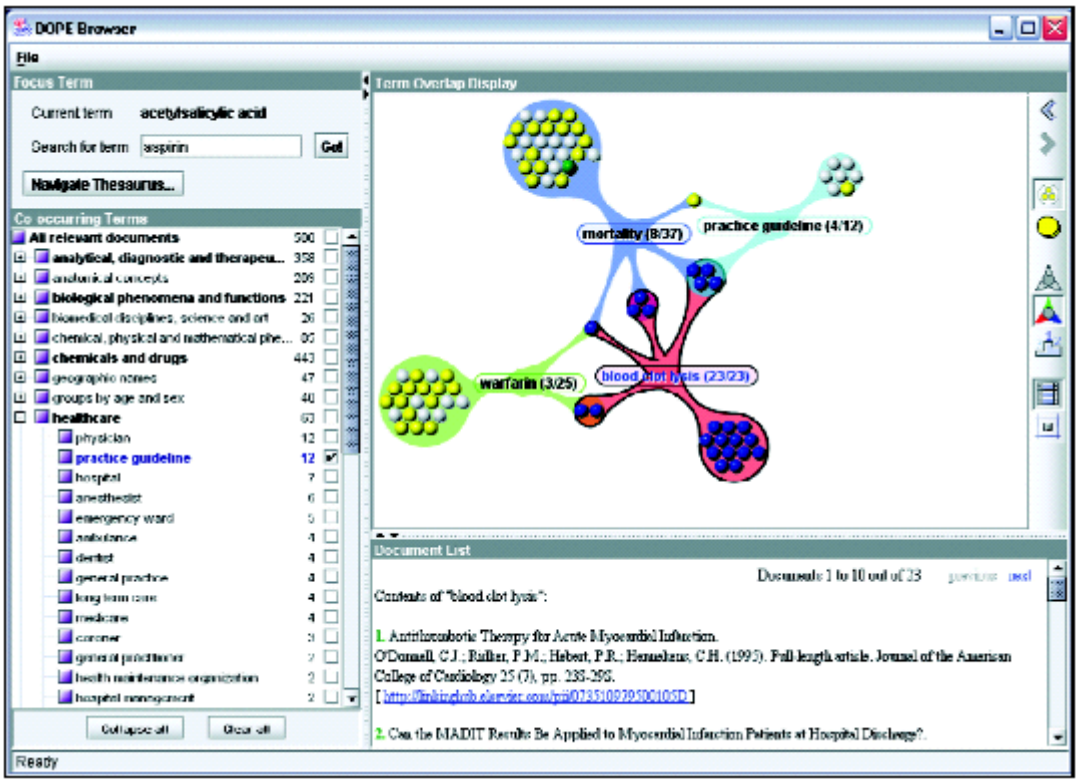


Figure 3.7 Interface d'exploration présentée dans [Stuckenschmidt 2004]

[Seeling 2003] propose un système dont le but est de permettre l'exploration de corpus à partir de l'analyse des méta-données associées aux documents ainsi que l'analyse de la similarité entre documents. Il s'appuie pour cela sur une ontologie répertoriant les méta-données utilisées dans le domaine de la finance. Cette ontologie est une hiérarchie de concepts. Les documents sont indexés manuellement à partir de cette ontologie. Le calcul de similarité entre les documents repose sur leurs méta-données communes. Une interface permet de visualiser d'une part l'ontologie des méta-données et d'accéder aux documents contenant ces méta-données. Elle est présentée dans la figure 3.8. D'autre part, l'interface représente les regroupements de documents dans un espace géo-spatial, l'utilisateur pouvant évaluer graphiquement la similarité entre les documents, sélectionner un document spécifique, visualiser les méta-données qui lui sont associées et accéder à son contenu.

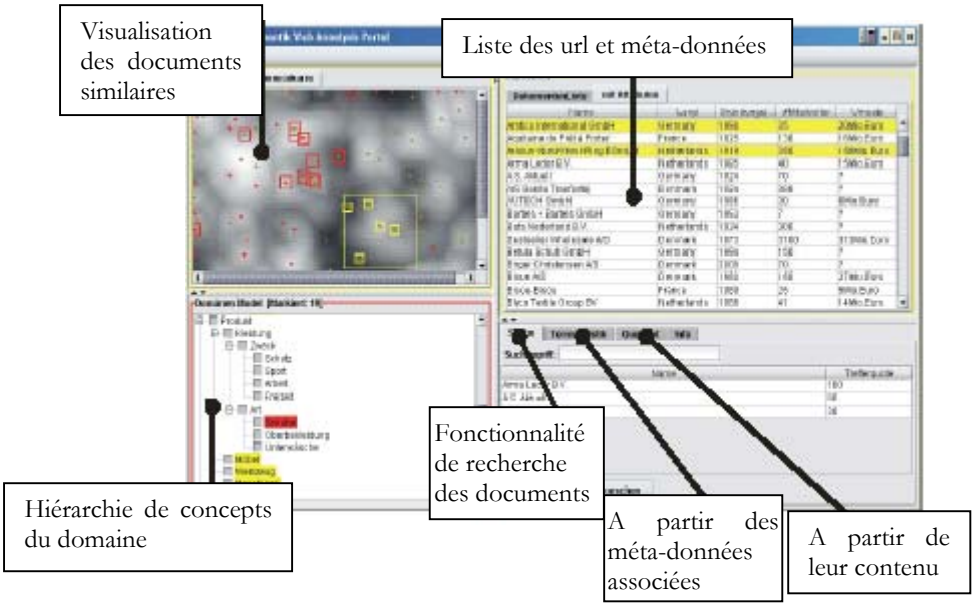


Figure 3.8 Interface d'exploration présentée dans [Seeling 2003]

L'intérêt de ces travaux est de prendre en compte à la fois les mots clés représentant le contenu des documents et les méta-données trouvées dans les documents. Cependant, la gestion (stockage et accès) des méta-données issues des documents n'est pas dissociée de la gestion des informations relatives au contenu. La réutilisabilité de ces systèmes sur de nouvelles collections du domaine paraît donc limitée, dans la mesure où de nouvelles méta-données ne seraient pas prises en compte par les systèmes. Une autre limite commune aux deux systèmes présentés précédemment est le manque de relations sémantiques entre les différents regroupements de documents construits pour l'exploration de la collection. L'utilisateur est certes intéressé par l'accès à des documents similaires et peut analyser les différents concepts du domaine influant sur ces regroupements ; cependant, il est regrettable de baser la représentation des documents sur des ontologies et de ne pas utiliser les relations sémantiques entre concepts pour présenter les relations entre documents à l'utilisateur.

5.4 Navigation dans un corpus à partir d'ontologies

Les portails communautaires sont un autre exemple d'applications visant à permettre le partage d'information et la communication entre membres de communautés à partir de techniques de navigation qui peuvent reposer sur des ontologies. Un portail a pour but de regrouper et de présenter des informations pertinentes pour la communauté ainsi que de permettre aux utilisateurs de publier des événements ou des informations. Un portail orienté Web Sémantique repose sur les technologies du Web Sémantique dans le but d'automatiser l'accès à l'information à travers des ontologies qui permettent une gestion de la sémantique des données et une communication entre les différents agents (machines et personnes). Les portails sémantiques sont décrits dans [Lausen 2004] suivant trois axes : les technologies sur lesquelles ils s'appuient, les procédés de traitement de l'information et l'accès à l'information. Nous nous concentrons sur ce dernier point qui est en relation avec notre domaine d'étude. A titre d'exemple, nous décrivons ici deux portails développés par des projets de recherche.

Le portail Esperonto[5], développé par l'Université Polytechnique de Madrid, vise à permettre la communication et le partage d'information entre les membres du projet européen

[5] www.esperonto.net

Esperonto. Une capture d'écran est présentée dans la figure 3.9. Le portail repose sur cinq ontologies formelles représentant les personnes, les documents, l'organisation, le projet et les rendez-vous. Ces ontologies ont des liens entre elles.

L'accès à l'information se fait par la navigation au sein de hiérarchies de concepts issues des cinq ontologies. La navigation est accessible à tout moment par la présence d'un cadre sur la partie gauche du site. Plusieurs fonctionnalités facilitent la navigation. La première fonctionnalité repose sur une interrogation par mots clés. Les mots clés sont recherchés dans la description ou le nom des différents éléments stockés sur le portail. La deuxième fonctionnalité consiste à naviguer au sein de l'ontologie et d'accéder à toutes les instances qui sont en lien avec le concept sélectionné. Le troisième accès repose sur une interface de visualisation plus élaborée qui permet à l'utilisateur de retrouver les concepts qui l'intéressent à partir de la spécification des propriétés que ces derniers doivent avoir. Cette fonctionnalité est en cours d'implantation.

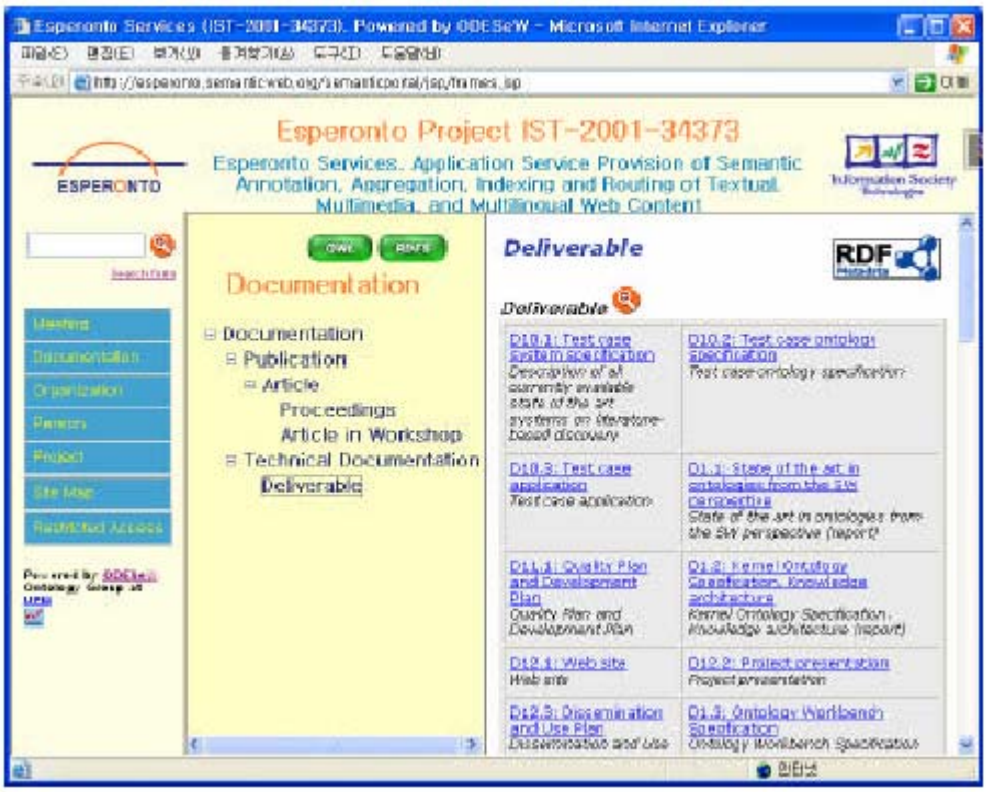


Figure 3.9 Copie d'écran du portail du projet Esperonto

Le portail OntoWeb[6] est un portail pour les chercheurs et industriels qui s'intéressent au Web Sémantique. Une capture d'écran est présentée dans la figure 3.10. Ce portail repose sur une ontologie formelle (en RDF) représentant l'organisation OntoWeb (entreprise, laboratoires de recherche), la documentation (delivrables, articles …), les événements et leur organisation. Il permet plusieurs types d'accès à l'information. Il contient deux menus de navigation. Le menu le plus général se trouve en haut des pages et permet l'accès aux différentes thématiques abordées dans le site. Le menu spécifique se trouve à la gauche de la page et permet de spécifier les ressources accessibles dans la thématique choisie.

[6] www.ontoWeb.org

Figure 3.10 Copie d'écran du portail OntoWeb

Deux fonctionnalités ont été implantées pour permettre un meilleur accès à l'information. La première fonctionnalité permet de naviguer dans l'ontologie et d'accéder aux instances associées aux concepts. A tout moment, l'utilisateur peut affiner ses recherches en spécifiant les propriétés et relations que les concepts ou instances doivent avoir avec le concept qu'il est en train de visualiser. Les éléments qui seront montrés à l'utilisateur seront ensuite filtrés par rapport à ces critères. La deuxième fonctionnalité est une recherche par mot clé. Cette recherche repose tout d'abord sur une indexation sémantique des ressources qui permet d'aller rechercher les ressources abordant les concepts spécifiés dans la requête. Dans le cas où aucune ressource n'est restituée, le système repose alors sur une indexation classique de la ressource et recherche l'occurrence des termes de la requête dans n'importe quelle partie de son contenu.

5.5 Bilan

L'utilisation d'ontologies pour la spécification du besoin utilisateur et la recherche des documents correspondant présente plusieurs avantages. Elle permet tout d'abord la reformulation de requêtes envoyées à des systèmes traditionnels à partir de l'ajout ou de la désambiguïsation de termes de l'ontologie. Dans le cas où la représentation des documents et de la requête repose sur une ontologie, elle permet également de mettre en place des mécanismes d'appariement faisant intervenir la similarité sémantique entre les concepts qui les composent. Les ontologies lourdes peuvent également mener à la mise en place de mécanismes d'inférences permettant la recherche des éléments de l'ontologie répondant à une requête à partir de langages d'interrogation sophistiqués.

En explorant une collection documentaire, l'utilisateur doit avoir une vue globale sur la collection à sa disposition et pouvoir naviguer au sein d'une représentation qui lui permet d'affiner sa recherche et d'accéder aux documents pertinents en fonction de son besoin. L'exploration à partir de hiérarchies de concepts a l'avantage de préciser le contexte de navigation. L'utilisateur a en effet accès à la thématique des classes à travers leur structuration hiérarchique. Il peut ainsi choisir de spécifier ou de généraliser sa recherche. L'exploration à partir d'ontologies permet d'enrichir le contexte d'exploration à partir de relations non taxonomiques entre les concepts. Des recherches avancées peuvent également être proposées pour accéder à

des informations élaborées. L'indexation conceptuelle permet également de pouvoir regrouper les documents à partir des concepts retrouvés dans le contenu et dans les méta-données. Nous pensons cependant, que ces deux types de connaissances, ontologies liées aux méta-données et ontologies liées au domaine traité, doivent être gérés de manière indépendante afin que le système puisse s'adapter à n'importe quelle collection. Des efforts doivent également être déployés pour permettre à l'utilisateur de situer son contexte d'exploration dans les ontologies. Les relations sémantiques reliant les concepts peuvent lui être présentées de façon intuitive pour ne pas surcharger la navigation.

6 Conclusion

Le choix de l'ontologie décrivant la connaissance sur laquelle un SRI peut reposer est, selon nous, une question non-triviale. Contrairement à certains auteurs pour qui le choix porte uniquement sur le domaine traité dans l'ontologie, nous considérons qu'une analyse quantitative de l'ontologie doit être réalisée afin d'établir si celle-ci contient la connaissance nécessaire à la compréhension du corpus. Les travaux de la littérature effectuent cette analyse en évaluant l'adéquation entre un corpus et une hiérarchie de concepts. Ils distinguent deux aspects dans l'analyse : l'évaluation de l'adéquation lexicale (à partir des termes extraits du corpus et des labels de l'ontologie) et l'évaluation de l'organisation des concepts dans la structure de l'ontologie. Cependant, les liens associatifs présents dans la structure d'une ontologie légère ne sont alors pas pris en compte.

Une fois l'ontologie choisie, son utilisation en RI a plusieurs objectifs. Les ontologies peuvent tout d'abord aider à l'interprétation du besoin utilisateur dans le cadre des SRI classiques. Par des mécanismes reposant sur l'ontologie, le sens de la requête initiale est alors spécifié soit par l'ajout de nouveaux termes en lien avec les concepts référés par les termes initiaux, soit par la désambiguïsation des termes. Les ontologies peuvent également servir à la représentation des granules documentaires et des requêtes à partir des concepts et instances associés à l'ontologie. Ceci correspond à l'indexation sémantique. L'indexation sémantique a pour avantage de représenter les documents à partir des éléments conceptuels auxquels ils font référence plutôt que par des termes bien souvent ambigus. Les mécanismes d'appariement entre le besoin utilisateur et la représentation des documents sont alors enrichis par la mise en correspondance de ces deux représentations à un niveau conceptuel. Des mesures de similarité entre les concepts de l'ontologie peuvent être utilisées pour évaluer soit le poids d'un concept dans un document, soit la similarité entre les concepts d'un document et ceux d'une requête. Cependant, la plupart des mesures de similarité de la littérature négligent les relations associatives présentes dans les ontologies légères qui permettent de spécifier la sémantique associée aux concepts. Les ontologies sont également utilisées pour faciliter l'accès aux informations contenues dans un corpus par des mécanismes d'exploration et de navigation. Afin d'aider l'utilisateur dans la spécification de son besoin en information, des vues sur la collection sont présentées à l'utilisateur à partir des concepts ayant servi à l'indexation des documents. Les mécanismes reposant sur des ontologies légères ou formelles ne dissocient pas les deux aspects liés à une activité de recherche qui sont le thème du corpus et les informations relatives à la tâche de recherche. Ceci mène à un faible potentiel de réutilisabilité des systèmes pour une autre tâche de recherche ou une collection abordant un autre thème. De plus, les techniques de navigation dans un corpus à partir d'ontologies présentent les concepts de celles-ci uniquement à partir de leurs relations taxonomiques, les relations associatives étant cachées à l'utilisateur. La spécification de ce dernier type de relations permettrait à l'utilisateur de mieux interpréter la sémantique qui est associée aux concepts.

Nos travaux visent à répondre aux différentes lacunes précédemment identifiées. Tout d'abord, nous proposons de différencier dans notre modélisation du contexte associée à une

recherche les aspects de tâche et de thème par l'utilisation d'ontologies légères de domaine dédiées à ces deux points. Le modèle est alors réutilisable pour différentes tâches de recherche et différents corpus. Nous proposons également d'intégrer ce modèle dans un SRI à deux niveaux. Premièrement, les granules documentaires sont indexés par rapport aux ontologies. Le mécanisme d'indexation sémantique proposé repose sur la proposition d'une mesure de similarité prenant en compte les relations associatives entre concepts. Ensuite, le modèle est pris en compte dans le SRI par la proposition d'un accès aux granules documentaires et aux informations qui leur sont associées à travers la navigation dans les deux ontologies. L'ensemble des relations établies pour un concept est présenté à l'utilisateur. Afin d'aider au choix de l'ontologie de domaine du thème traité dans le corpus, nous proposons également de définir une méthodologie et des méthodes pour évaluer l'adéquation entre un corpus et une ontologie légère en considérant l'ensemble de ces éléments (lexique, structure : relations taxonomiques et associatives).

L'ensemble de ces différents points sont décrits dans les chapitres suivants.

Partie 2
Contributions

Chapitre 4
Modèle

1 Introduction

Dans ce chapitre, nous envisageons la modélisation du contexte en RI selon deux angles complémentaires : le thème et la tâche dans lesquels s'inscrit le besoin en information.

Dans la littérature, le thème du besoin en information de l'utilisateur est considéré uniquement au travers des mots clés apparaissant dans la requête. La sélection des documents repose sur un appariement entre les termes de la requête et ceux des documents. Ce principe pose différents types de problèmes. L'ambiguïté des termes peut conduire à la sélection de documents non pertinents. De plus, la prise en compte de la terminologie plutôt que de la sémantique peut conduire à la non sélection de documents pertinents, dans le cas où un même concept est référencé par deux termes différents dans la requête et dans les documents. Les techniques de reformulation de requêtes comme la réinjection de pertinence [Harman 1992] ont pour objectif de traiter ces problèmes en rajoutant à la requête de nouveaux termes qui sont issus de documents pertinents. Il s'agit là d'un premier pas vers la prise en compte de la sémantique issue des documents. Cependant, ce traitement reste au niveau terminologique et peut ne pas résoudre les problèmes d'ambiguïté, il peut même en rajouter.

L'indexation sémantique vise à représenter les granules documentaires (documents, parties de document, ou ensembles de documents) non plus au niveau terminologique mais au niveau conceptuel, c'est-à-dire au niveau des objets ou concepts référencés par les termes des granules [Haav 2001] [Guha 2003] [Kiryakov 2004]. Dans ce cadre, l'utilisation d'une ontologie générique, telle que WordNet, est adaptée à des corpus couvrant de multiples domaines, comme par exemple les nouvelles issues de journaux [Khan 2002] [Cucchiarelli 2004]. Elle pose cependant le problème de la couverture terminologique : seuls les termes les plus communs sont référencés. Une telle indexation doit donc être couplée à une indexation terminologique traditionnelle (termes extraits des granules) [Gonzalo 1998] [Hernandez 2003a] [Baziz 2005]. Un autre problème est la complexité de l'étape de désambiguïsation. Les termes polysémiques, nombreux dans ce type de ressources, référencent plusieurs concepts. Le choix du concept à associer aux granules est donc problématique. L'utilisation d'ontologies de domaine limite ces deux problèmes. Etant focalisée, la terminologie complète peut être intégrée à l'ontologie. De plus, le nombre de sens par termes est plus restreint [Furst 2004]. Dans les approches existantes, l'indexation sémantique s'appuie sur des ressources comportant essentiellement des liens hiérarchiques entre concepts (plus spécifiques, plus génériques). La désambiguïsation repose sur ce type de lien pour appréhender le sens d'un terme. Or, l'utilisation d'autres liens sémantiques pourrait améliorer ces mécanismes.

Ainsi, le modèle que nous proposons repose sur des ontologies de domaine, pour structurer la connaissance du domaine du thème traité dans le corpus. Les ontologies sur lesquelles nous nous appuyons intègrent des liens hiérarchiques ainsi que des liens associatifs (tels que « est une partie de », « est un phénomène lié à », …).

Outre le domaine du thème, nous nous intéressons également à la tâche dans laquelle s'inscrit la recherche de l'utilisateur. La tâche se réfère à l'activité pour laquelle la recherche d'information est effectuée. Par exemple, dans le cadre d'une tâche de veille, l'utilisateur recherche des granules pour connaître les acteurs importants, l'évolution du domaine, etc. Dans le cadre d'une tâche d'apprentissage pédagogique, en revanche, l'utilisateur voudra retrouver une illustration d'une définition, un exercice en rapport avec la notion étudiée, etc. Les modèles actuels de RI intègrent peu cet aspect. Lorsque cet aspect est modélisé, il est confondu dans la modélisation du domaine du thème. Les méta-données relatives à la tâche (auteur, type de document, etc.) sont des concepts de l'ontologie légère au même titre que les concepts associés aux thèmes [Kiryakov 2004] [Vallet 2005]. Cette approche empêche la réutilisation du modèle pour différentes tâches.

Notre modèle, au contraire, sépare les aspects de tâche et de thème tout en les mettant en relation. Chaque aspect est modélisé par une ontologie de domaine légère. Cette formalisation permet d'établir les connaissances associées à ces deux aspects à travers des relations sémantiquement riches. Elle permet en outre de déduire de nouvelles connaissances à partir du lien entre ces ontologies. Ce modèle est intégré dans un processus de RI à travers deux mécanismes : la représentation des granules et l'accès à l'information. La représentation des granules repose sur des travaux de la littérature. Cependant, concernant l'indexation sémantique (lien entre l'ontologie de thème et les granules), ces travaux ne prennent en compte que les liens hiérarchiques entre concepts. Nous proposons une extension qui intègre les relations associatives dans le calcul du pouvoir discriminant d'un concept. Concernant l'annotation (lien entre l'ontologie de tâche et les granules), nous nous appuyons sur des techniques d'extraction d'instances à partir des méta-données contenues dans les granules. La formalisation de la connaissance à travers deux ontologies légères permet un accès à l'information plus élaboré qu'au travers de requêtes en langage libre. La thématique est en effet enrichie par la prise en compte de la tâche. Un concept donné n'aura pas le même sens en fonction de l'élément de tâche auquel il est associé. Par exemple, dans le cas de la tâche de veille, le concept *« X ray »* peut intéresser l'utilisateur soit par rapport aux publications qui contiennent ce concept, soit par rapport aux chercheurs dont il est le domaine d'intérêt. Ainsi, nous proposons un accès à l'information qui combine thème et tâche. Cette formalisation permet également une interrogation à deux niveaux d'abstraction : au niveau des concepts, l'utilisateur peut avoir une vue globale du contenu de l'information ; au niveau instance, l'utilisateur accède aux données elles-mêmes, qu'elles soient des granules, des méta-données ou de l'information élaborée.

Le modèle est présenté dans les deux sections suivantes. La section 2 décrit notre modèle du contexte sémantique en abordant d'abord la modélisation du domaine de la tâche, puis celui du thème traité dans le corpus ainsi que le lien entre ces deux ontologies. La section 3 présente l'intégration du modèle au processus de RI aux deux niveaux indiqués plus haut : la représentation des granules et l'accès à l'information.

2 Modélisation du contexte sémantique

Dans notre modèle, nous proposons de prendre en compte deux facteurs dont dépend un besoin en information : le thème du besoin et la tâche pour laquelle l'utilisateur effectue une recherche [Hernandez 2005a].

Le thème d'un besoin en information doit être mis en relation avec les thématiques associées aux domaines des connaissances couverts par le corpus. L'utilisateur spécifie en effet son besoin à partir des informations qu'il pense pouvoir retrouver dans la collection. Pour qu'il puisse préciser son besoin en fonction des thématiques effectivement traitées dans le corpus, il est nécessaire que celles-ci soient situées dans le contexte général du ou des domaines abordés. Le contexte doit alors rendre compte des thématiques liées au(x) domaine(s) de connaissance(s) abordé(s) dans le corpus et permettre de situer les thématiques du corpus dans ce cadre.

Un utilisateur effectue une tâche de recherche pour accéder à des données précises répondant à son besoin. Le contexte doit donc également permettre de rendre explicites les informations recherchées ainsi que les raisons pour lesquelles elles le sont et comment elles seront utilisées. Dans notre modèle, ceci est réalisé au travers des méta-données intéressant l'utilisateur et qui sont associées aux granules. Les méta-données et leurs liens, implicites ou explicites, spécifient les objets de la recherche ainsi que leur rôle dans la tâche. Par exemple, dans le cadre d'une exploration de corpus dans le but d'effectuer de la veille sur les activités de recherche d'un laboratoire, les informations intéressant l'utilisateur peuvent être les chercheurs du laboratoire, les types de publications, les années des publications, les thématiques de recherche

ainsi que les corrélations entre ces différents acteurs (thématiques abordées pendant les cinq dernières années, chercheur ayant le plus publié…). A travers les méta-données du corpus, ces informations peuvent être capturées et restituées à l'utilisateur.

Pour modéliser ces deux aspects du contexte d'une recherche, des ontologies légères de domaine sont utilisées [Hernandez 2003b]. L'intérêt principal d'utiliser des ontologies est de préciser le contexte *sémantique* lié à la recherche. Les thématiques abordées dans le granule et les méta-données qui en sont extraites sont alors représentées par « une spécification explicite et formelle d'une conceptualisation partagée » [Studer 1998]. Le contexte est modélisé en terme de concepts et de relations sémantiques liés à ces différents aspects. Ceci permet en premier lieu de préciser le contexte de façon formelle et non ambiguë, et par conséquent il peut être interprété par le système et l'utilisateur à travers la sémantique représentée. Une ontologie fournit en effet une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine [Aussenac 2004]. De plus, le contexte peut être intégré au SRI sur la base de nouvelles connaissances déduites de celles représentées dans les ontologies. Le traitement des informations contenues dans le corpus bénéficie de ces mécanismes élaborés.

Les ontologies utilisées dans notre modèle doivent permettre la représentation du contexte à la fois à travers le domaine de connaissances abordé dans le corpus, mais également au travers de l'ensemble des méta-données qui peuvent être utiles dans des activités d'accès à l'information et d'exploration. Contrairement aux approches existantes qui intègrent dans une même ontologie les thématiques d'un domaine et les méta-données associées au granule [Benjamins 1999] [Kiryakov 2004] [Vallet 2005], le modèle proposé vise à séparer ces deux aspects afin de s'adapter à n'importe quel domaine et n'importe quelle tâche. Plusieurs tâches peuvent en effet être effectuées sur un même corpus en fonction des données recherchées par l'utilisateur. Les thématiques du corpus restent les mêmes mais leur intégration dans la recherche dépend alors du contexte lié à la tâche à accomplir. De la même façon, une même tâche peut être effectuée sur un corpus d'un autre domaine, les acteurs de la recherche (les chercheurs ou les publications en reprenant notre exemple précédent) restent identiques mais leurs corrélations au domaine traité dans le corpus sont modifiées en fonction de celui-ci. Dans un souci d'adaptabilité, le modèle proposé repose donc sur deux ontologies différentes : une ontologie du domaine de la tâche à réaliser et une ontologie du domaine abordé dans le corpus. Ces deux types d'ontologie font appel à des concepts différents qui sont décrits dans les sections 2.1 et 2.2.

2.1 Formalisation

2.1.1 Ontologie légère de domaine

Les ontologies considérées sont des ontologies dites légères. Ces ontologies sont décrites à partir d'un lexique L et d'une structure S [Maedche 2002b]. Le niveau lexical couvre tous les termes ou labels définis pour désigner les concepts et les relations. Le niveau conceptuel défini dans la structure de l'ontologie représente les concepts et la sémantique qui leur est associée à partir des relations conceptuelles qui les lient. Nous considérons de plus que, dans ces ontologies, les caractéristiques des relations associatives peuvent être précisées (synonymie, transitivité,…).

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)
- σ_R : R $\rightarrow$ C x C est la signature d'une relation associative
- σ_A : A $\rightarrow$ C x T est la signature d'une relation d'attribut
- σ_{CARR} : R$\rightarrow$ 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.

2.1.2 Modèle du domaine

Le modèle d'un domaine est représenté à partir d'une ontologie, telle que présentée dans la section précédente et des instances qui sont associées à ses concepts.

Il se formalise par le tuple suivant (O,I, V, f_C, f_T, f_R, f_A) avec :

- O l'ontologie légère du domaine
- I l'ensemble des instances
- V l'ensemble des valeurs des types de données
- f_C : I $\rightarrow$ C est la fonction d'instanciation d'un concept
- f_T: V $\rightarrow$ T est la fonction d'instanciation d'un type de donnée
- f_R : I x I $\rightarrow$ R est la fonction d'instanciation d'une relation associative
- f_A : I x V $\rightarrow$ A est la fonction d'instanciation d'une relation d'attribut

2.2 Ontologie du domaine de la tâche

2.2.1 Définitions

2.2.1.1 Ontologie

Une **ontologie du domaine de la tâche** est constituée d'un ensemble de concepts intéressant l'utilisateur dans la tâche de RI qu'il entreprend et d'un ensemble de relations spécifiant leurs rôles dans la tâche. L'intérêt d'utiliser cette ontologie pour préciser le contexte de la tâche est que les relations sémantiques définies permettent de spécifier le rôle des concepts de tâche entre eux. Elle est choisie pour une tâche donnée par un utilisateur ou un groupe

d'utilisateurs en fonction des données ciblées par la recherche. Elle peut être soit réutilisée, soit élaborée pour la tâche en question, l'essentiel étant qu'elle soit validée par l'utilisateur et que son lexique et sa structure spécifient effectivement la conceptualisation liée à la tâche qu'il souhaite réaliser. L'ontologie du domaine de la tâche est réutilisable pour la même tâche sur différents corpus.

Cette ontologie du domaine de la tâche est donc constituée :

- d'une structure $S_{tâche} := \{ C_{tâche}, R_{tâche}, A_{tâche}, T_{tâche}, CAR_{Rtâche}, \leq^{C}_{tâche}, \sigma_{Rtâche}, \sigma_{Atâche}, \sigma_{CARR} \}$
- d'un lexique $L_{tâche} := \{L^{C}_{tâche}, L^{R}_{tâche}, F_{tâche}, G_{tâche}, \}$

Dans la majorité des tâches (système de question-réponse, veille, RI sur les gènes, …), les concepts intéressant l'utilisateur sont des méta-données associées explicitement ou implicitement aux documents. Nous nous focalisons sur ce type de tâche. Par abus de langage, nous utiliserons le terme « ontologie de tâche » pour désigner l'« ontologie du domaine de la tâche ».

2.2.1.2 Modèle du domaine de la tâche

Le modèle du domaine de la tâche est défini à partir de l'ontologie de la tâche présentée précédemment, et des instances de ses concepts extraits des documents ainsi que des fonctions permettant d'instancier les éléments de l'ontologie. Le modèle est donc constitué de $\{O_{tâche}, I_{tâche}, V_{tâche}, f_{Ctâche}, f_{Ttâche}, f_{Rtâche}, f_{Atâche}\}$

Le modèle est propre au corpus pour lequel il spécifie le contexte de la tâche. Les techniques que nous verrons dans la section 3.1 permettent d'extraire ces instances.

2.2.2 Illustration 1 : cas de la veille

La tâche de veille est une activité par laquelle un expert observe et analyse l'environnement scientifique, technique et technologique d'un domaine pour en déduire par exemple les opportunités de développement ou analyser la concurrence. Dans le cadre du Projet Masse de Données en Astronomie[1] (décrit dans le chapitre 7), l'activité de veille considérée consiste à analyser l'évolution des thématiques de recherche étudiées ainsi que les différents acteurs intervenant dans ces recherches (laboratoire, chercheur, pays, …).

Une ontologie référant aux activités de veille a été construite manuellement en coopération avec des astronomes à partir de la méthodologie Methontology décrite dans la section 2.2.1 du chapitre 2. La conception repose sur les travaux précédemment réalisés dans [Mothe 2004]. Cette ontologie a pour but de permettre une cartographie du domaine en présentant les acteurs du domaine. Les chercheurs auteurs d'articles, leur laboratoire d'appartenance, leur pays d'affiliation, ainsi que leurs articles sont représentés. D'autres informations sont mémorisées telles que l'aspect temporel, les journaux du domaine dans lesquels les articles scientifiques sont publiés. Enfin, les objets sur lesquels portent les articles et qui correspondent aux domaines d'intérêt des chercheurs correspondent à un concept particulier qui fait le lien avec l'ontologie de thème et que nous verrons dans la section 2.4.

Le modèle du domaine de la veille associé est construit automatiquement à partir des méta-données connues et utiles pour la tâche visée. C'est-à-dire que les concepts et leurs relations sont choisis a priori, mais des techniques automatiques permettent d'extraire du corpus leurs

[1] http://cdsWeb.u-strasbg.fr/MDA/mda.html

instances et les valeurs des propriétés d'attributs de façon automatique. Ces techniques d'extraction sont décrites dans la section 3.1.2.

La structure du modèle de la tâche est $\{I_{tâche}, C_{tâche}, R_{tâche}, A_{tâche}, T_{tâche}, V_{tâche}, CAR_{Rtâche}, \leq^{C}_{tâche}, \sigma_{Rtâche}, \sigma_{CARRtâche}, \sigma_{Atâche}, f_{Ctâche}, f_{Ttâche}, f_{Rtâche}, f_{Atâche}\}$ avec

- $C_{tâche}$ = {Chercheur, Laboratoire, Article, Pays, Littérature, Date, Objet, Revue, Ouvrage, Actes}
- $R_{tâche}$ = {co-ecrit, travaille_pour, a_pour_dom_d'intérêt, rédige, ecrit_en, traite_de, publié_dans, est_situé_en}
- $A_{tâche}$ = {a_pour_nom, a_pour_prenom, a_pour_référence, a_pour_adresse}
- $I_{tâche}$ = { $chercheur_1$,…,$chercheur_n$, $laboratoire_1$,…, $laboratoire_n$, $article_1$,…, $article_n$, $pays_1$,…,$pays_n$, $date_1$,…,$date_n$, $objet_1$,…, $objet_n$, $revue_1$, …, $revue_n$, … }
- $T_{tâche}$ = {String }
- $CAR_{Rtâche}$ = {symétrie, transitivité, fonctionnalité}
- $V_{tâche}$ = {Dupond, Jean, A-1, A-2, 118 route de Narbonne 31400 Toulouse,…}
- $\leq^{C}_{tâche}$ = {(Revue, Littérature),(Ouvrage, Littérature), (Actes, Littérature)}
- $\sigma_{Rtâche}$ = { (co-ecrit (Chercheur, Chercheur)), (travaille_pour (Chercheur, Laboratoire)), rédige(Chercheur, Article), ecrit_en(Article, Date), a_pour_dom_d'intérêt(Chercheur, Objet), traite_de(Article, Objet), publié_dans(Article, Litterature), est_situé_en(Laboratoire, Pays)}
- $\sigma_{CARRtâche}$ = {(co-ecrit,symétrie)}
- $\sigma_{Atâche}$ = {(a_pour_nom(Chercheur, String),(a_pour_prenom(Chercheur,String)), (a_pour_référence(Article, String)), (a_pour_adresse(Laboratoire,String))}
- $f_{Ctâche}$ = {(chercheur1, Chercheur1), (laboratoire1, Laboratoire), …}
- $f_{Ttâche}$ = {(Dupond, String), (Jean, String), (A-1, String), …}
- $f_{Rtâche}$ = {(co-ecrit (chercheur1, chercheur2)), (travaille_pour (chercheur1, laboratoire1)), (rédige(chercheur1, article1)), …}
- $f_{Atâche}$ = {(a_pour_nom(chercheur1, Dupond)), (a_pour_prenom(chercheur1, Jean)), (a_pour_référence(article1, A-1)), …}

Le lexique $L_{tâche}$:= $\{L^{C}_{tâche}, L^{R}_{tâche}, F_{tâche}, G_{tâche}\}$

- $L^{C}_{tâche}$ = {« chercheur », « laboratoire », « institut de recherche », …}
- $L^{R}_{tâche}$ = {« travaille pour », « rédige », « est situé en », « a pour domaine d'intérêt », …}
- $F_{tâche}$ = {(Chercheur, « chercheur »), (Laboratoire, « laboratoire »), (Laboratoire, institut de recherche »), …}
- $G_{tâche}$ = {(travaille_pour, « travaille pour ») (est_situé_en, « est situé en), …}

Une représentation schématique au niveau de l'ontologie de tâche est présentée figure 4.1.

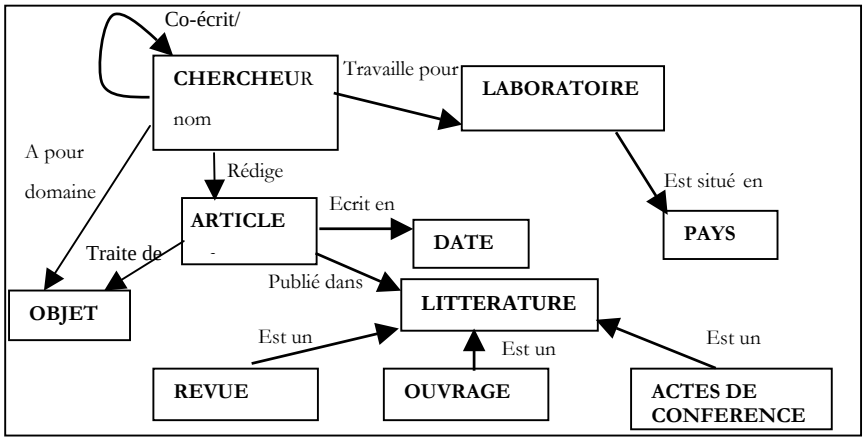


Figure 4.1 Ontologie du domaine de la veille

2.2.3 Illustration 2 : cas de l'apprentissage pédagogique

Une ontologie se rapportant à une tâche d'apprentissage pédagogique a été réalisée en co-opération avec des membres de l'Institut Universitaire de Formation des Maîtres de Midi Pyrénées. Elle a pour but de préciser les contextes de l'apprentissage en spécifiant les ressources disponibles (ouvrage, logiciel, …), les modules qui composent ces ressources, leur type (cours, exercices, évaluation) ainsi que l'ordre dans lequel ils doivent être étudiés (relation « suit »). Une représentation schématique au niveau des concepts de cette ontologie de tâche est présentée figure 4.2. D'autres ontologies, telles que celle présentée dans [Angelova 2004], pourraient être utilisées. L'ontologie présentée dans cette section répond cependant à des besoins spécifiques qui ont été émis au moment de sa création.

Le modèle du domaine de l'apprentissage pédagogique est formalisé par :

Une structure $S_{tâche}=\{I_{tâche}, C_{tâche}, R_{tâche}, A_{tâche}, T_{tâche}, V_{tâche}, CAR_{Rtâche}, \leq^{C}_{tâche}, \sigma_{Rtâche}, \sigma_{CARRtâche}, \sigma_{Atâche}, f_{Ctâche}, f_{Ttâche}, f_{Rtâche}, f_{Atâche}\}$ avec

- $C_{tâche} = \{$Module, Evaluation, Exercice, Cours, Ressource, Manuel, Logiciel, Objet$\}$
- $R_{tâche}$ = {contient, aborde_la_notion, est_mis_en_pratique_dans, suit, utilisé_dans}
- $A_{tâche}$ = {a_pour_référence }
- $I_{tâche}=\{$ module1_1,…,module$_n$, exercice$_1$,…exercice$_n$, ressource$_1$… $\}$
- $T_{tâche} = \{$String $\}$
- $V_{tâche} = \{$C1, C2, M1,…$\}$
- $\leq^{C}_{tâche} = \{($ Logiciel, Ressource),(Manuel, Ressource) $\}$
- $\sigma_{Rtâche} = \{$ (contient (Ressource, Module)), (aborde_la_notion (Module, Objet)), (est_mis_en_pratique_dans (Cours, Exercice)), … $\}$
- $\sigma_{Atâche}=\{$(a_pour_référence(Module, String),(a_pour_référence (Cours,String)), (a_pour_référence(Exercice, String)) …$\}$
- $f_{Ctâche}=\{$(exercice1, Exercice), (module1, Module), …$\}$
- $f_{Ttâche}=\{$(C1, String), (C2, String), (M1, String), …$\}$

- $f_{Rtâche}$={(contient (ressource$_1$, module$_1$)), (est_mis_en_pratique_dans (cours1, exercice1)) …}
- $f_{Atâche}$={(a_pour_référence(module1, M1)), (a_pour_référence(cours1, C1)), …}

Le lexique $L_{tâche}$:={$L^C_{tâche}$,$L^R_{tâche}$, $F_{tâche}$,$G_{tâche}$}

- $L^C_{tâche}$,={ « module », « evaluation », « exercice », « cours », « ressource », « manuel », « logiciel », « objet » }
- $L^R_{tâche}$ ={ « contient », « aborde la notion », « est mis en pratique dans », « suit », « utilisé dans » …}
- $F_{tâche}$ ={(module, « module »), (cours, « cours »), …}
- $G_{tâche}$ ={(aborde_la_notion, « aborde la notion ») (utilisé_dans, « utilisé dans »), …}

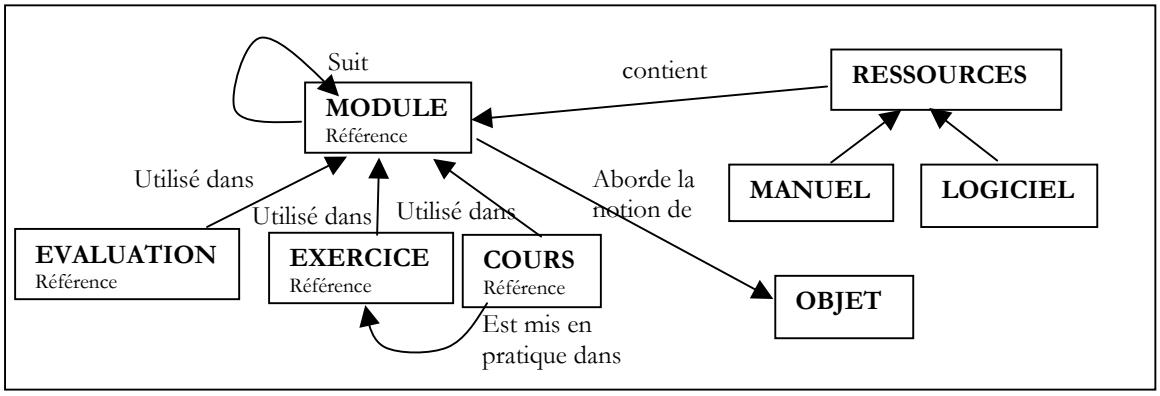


Figure 4.2. Ontologie du domaine de l'apprentissage pédagogique

2.3 Ontologie du domaine du thème traité dans le corpus

L'**ontologie du domaine du thème abordé dans le corpus** représente le domaine traité dans le contenu des documents du corpus. Elle représente, à travers son lexique et sa structure, la connaissance liée au thème abordé dans le corpus. Par l'accès à cette ontologie, l'utilisateur pourra situer son expertise par rapport au domaine modélisé, il pourra ensuite spécifier son besoin à partir de l'interprétation qu'il fera du contexte.

Cette ontologie du thème est donc constituée :

- d'une structure $S_{thème}$:={$C_{thème}$, $R_{thème}$, $A_{thème}$, $T_{thème}$, $CAR_{Rthème}$, $\leq^C_{thème}$, $\sigma_{Rthème}$, $\sigma_{Athème}$, $\sigma_{CARRthème}$ }
- d'un lexique $L_{thème}$:={$L^C_{thème}$,$L^R_{thème}$, $F_{thème}$,$G_{thème}$ }

Le domaine du thème que nous traitons ne comporte pas d'instance. L'ontologie suffit et le modèle du domaine du thème n'est pas nécessaire. Comme nous le soulignons dans les différentes sections présentant les cas d'étude auxquels nous appliquons notre modèle, les instances et relations d'attributs peuvent ne pas être nécessaires dans la modélisation du contexte associé au thème. La modélisation pourrait cependant être étendue pour prendre en compte ces éléments en considérant le modèle du thème, dans le cas où celui-ci aurait un intérêt pour les futurs utilisateurs du SRI.

La modélisation du contexte se veut générique et peut s'adapter à n'importe quelle ontologie liée au domaine visé. Cette ontologie peut être élaborée à cet effet par exemple par la transformation d'un thésaurus en ontologie (cf chapitre 5). Elle peut également être réutilisée alors qu'elle a été créée dans d'autres perspectives. Cependant, comme nous le soulignons dans le chapitre 6, il est indispensable que le choix de l'ontologie prenne en compte son adéquation au corpus [Hernandez 2004b].

Par abus de langage, l'« ontologie du domaine du thème » est appelée « ontologie du thème ».

2.3.1 Illustration 1 : cas du domaine de l'astronomie

Dans le cadre des travaux présentés dans cette thèse, cette ontologie s'attache à définir la connaissance liée au domaine de l'astronomie (objets astronomiques, techniques, phénomènes physiques, théories). Une partie de l'ontologie de l'astronomie utilisée est représentée dans la figure 4.3. Notons que dans le domaine de l'astronomie différents outils (tels que Simbad[2] intégré au serveur bibliographique ADS[3] voir chapitre 7) ont été développés pour permettre l'accès aux différentes instances d'objets, phénomènes ou théories (astre ou phénomène donnés) présents dans un corpus. La modélisation du domaine du thème intéressant les acteurs de ce domaine se situe donc plutôt au niveau conceptuel qu'au niveau des concepts car elle permet l'intégration d'un nouveau type de connaissance. La prise en compte du modèle du thème pourrait toutefois être envisagée par l'intégration à notre modèle des instances issues de bases de données du domaine.

La structure de l'extrait de l'ontologie est $\{C_{thème}, R_{thème}, \leq^{C}_{thème}, \sigma_{Rthème}, f_{Rthème},\}$ avec

- $C_{thème}$ ={Corps_céleste, Astéroïde, Comète, Système_solaire, Etoile, Soleil, Eclipse_solaire, Couronne_solaire, Objet_astronomique}
- $R_{thème}$ = {est_une_partie_de, est_un_événement_lié_à }
- $\leq^{C}_{thème}$ ={ (Corps_céleste, Objet_astronomique), (Comète, Corps_céleste), (Astéroïde, Corps_céleste), (Système_solaire, Corps_céleste), (Soleil, Etoile)}
- $\sigma_{Rthème}$ ={ (est_une_partie_de(Comète, Système_solaire)), (est_une_partie_de (Astéroïde, Système_solaire)), est_une_partie_de (Soleil, Système_solaire), est_une_partie_de (Couronne_solaire, Soleil), est_un_événement_lié_à (Eclipse_solaire, Soleil) }

Le lexique $L_{thème}$:={$L^{C}_{thème}$,$L^{R}_{thème}$, $F_{thème}$,$G_{thème}$}

- $L^{C}_{thème}$={«corps céleste », « comète », « astéroïde », « planétoïde », « système solaire » , « étoile », « soleil », « éclipse solaire », « couronne solaire » …}
- $L^{R}_{thème}$ ={« est une partie de », « est un événement lié à »}
- $F_{thème}$ ={(Astéroïde, « astéroïde »), (Astéroïde, « planétoïde »), (Corps_céleste, « corps céleste »), (Système_solaire, « système solaire »…}
- $G_{thème}$ ={(est_une_partie_de, « est une partie de ») (est_un_événement_lié_à, « est un événement lié à »), …}

---

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

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

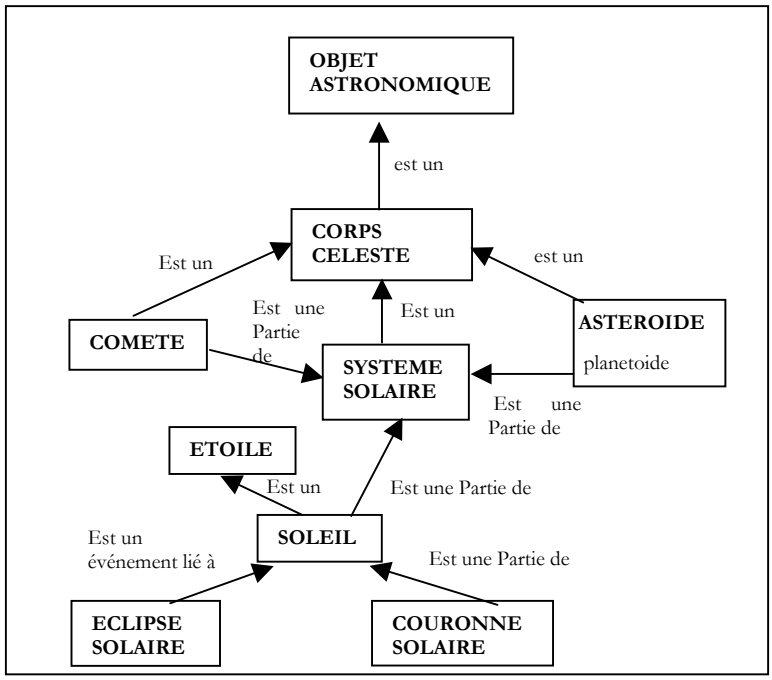


Figure 4.3 Extrait d'une ontologie du domaine du thème de l'astronomie

2.3.2 Illustration 2 : cas du domaine du référentiel du BTS informatique

Une ontologie relative au thème abordé dans le cadre du référentiel du BTS informatique a été proposée. Cette ontologie spécifie les notions qui doivent être assimilées par des étudiants de ce type de formation. Un extrait de cette ontologie est présenté figure 4.4. La modélisation du domaine du thème se limite à une ontologie et n'intègre pas les instances du domaine. Les objectifs pédagogiques de la formation visent à présenter le domaine de l'informatique à travers ces différentes notions et objets. Ils n'ont pas pour but de former les étudiants sur des considérations techniques propres aux instances des concepts modélisés telle que la particularité d'un processeur Intel Centrino ou Pentium 4.

La structure de l'extrait de l'ontologie est $\{C_{thème}, R_{thème}, \leq^{C}_{thème}, \sigma_{Rthème}, f_{Rthème},\}$ avec

- $C_{thème}$ = {Savoir_BTS_informatique, Architecture_des_ordinateurs, Développement_d_applications, Composant, Architecture_des_réseaux, Processeur, Mémoire}
- $R_{thème}$ = {est_une_partie_de}
- $\leq^{C}_{thème}$ ={(Processeur, Composant),(Mémoire, Composant), (Architecture_des_ordinateurs, Savoir_BTS_informatique), (Développement_d_applications, Savoir_BTS_informatique)}
- $\sigma_{Rthème}$ ={ (est_une_partie_de(Composant, Architecture_des_ordinateurs)), (est_une_partie_de (Architecture_des_réseaux, Architecture_des_ordinateurs)) }

Le lexique $L_{thème}$:={$L^{C}_{thème}$,$L^{R}_{thème}$, $F_{thème}$,$G_{thème}$}

- $L^{C}_{thème}$={ « savoir BTS informatique », « architecture des ordinateurs », « développement d'applications », « composant », « architecture des réseaux », « processeur », « mémoire » }
- $L^{R}_{thème}$ ={« est une partie de » }

- $F_{thème}$ ={(Savoir_BTS_informatique, « savoir BTS informatique », (Architecture_des_ordinateurs, « Architecture des ordinateurs »), …}
- $G_{thème}$ ={(est_une_partie_de, « est une partie de ») }

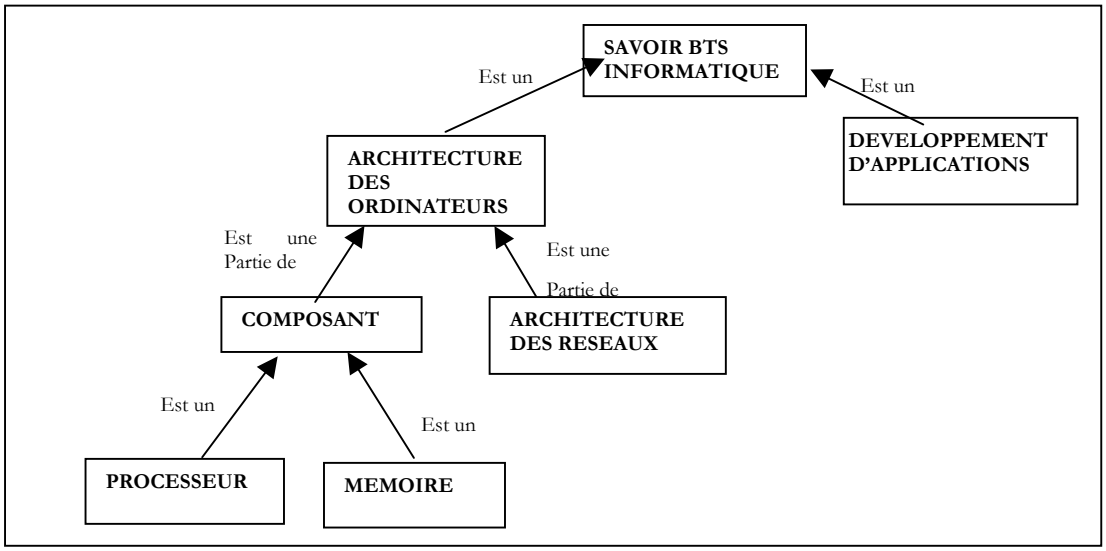


Figure 4.4 Extrait d'une ontologie du domaine du thème du référentiel BTS en informatique

2.4 Liens entre les deux ontologies

Les deux ontologies sont nécessaires pour réaliser une tâche de RI. Il est donc important que ces deux ontologies distinctes soient accessibles l'une par l'autre. Les concepts de l'ontologie de thème peuvent ainsi être rattachés aux concepts de l'ontologie de tâche, permettant de les situer dans le contexte de la tâche. Les objets du thème introduits dans l'ontologie de la tâche peuvent couvrir tout le domaine du thème. Il est également intéressant de prendre en compte la connaissance établie à partir des deux ontologies : la connaissance qu'elles représentent peut être combinée pour que de nouvelles connaissances soient déduites de ces deux représentations. Des relations entre concepts de l'ontologie de tâche pourront être déduites à partir du lien entre les ontologies.

Le lien est établi au travers du concept *Objet* présent dans chacune des ontologies de tâche. Il correspond à l'*Objet commun*. Ses instances ou valeurs sont prises par les concepts de l'ontologie de thème. Ainsi, l'ontologie de tâche définit en quelque sorte les rôles des concepts de l'ontologie du thème par rapport à la tâche.

Soient la structure du modèle du domaine de la tâche et la structure de l'ontologie du thème :

$S_{tâche}$ {$I_{tâche}$, $C_{tâche}$, $R_{tâche}$, $A_{tâche}$, $T_{tâche}$, $V_{tâche}$, $\leq^{C}_{tâche}$, $\sigma_{Rtâche}$, $\sigma_{Atâche}$, $f_{Ctâche}$, $f_{Ttâche}$, $f_{Rtâche}$, $f_{Atâche}$ } et $S_{thème}$ { $C_{thème}$, $R_{thème}$, $A_{thème}$, $T_{thème}$, $\leq^{C}_{thème}$, $\sigma_{Rthème}$, $\sigma_{thème}$}.

Le concept *Objet_commun* appartient à $C_{tâche}$.

La fonction d'instanciation associée à ce concept appartenant à $f_{Ctâche}$ est :

f_c(Objet_commun,i), tel que $i \in C_{thème}$.

Notons que la modélisation que nous proposons pour le domaine du thème n'intègre pas d'instance. Ce choix est lié aux différents cas d'étude que nous avons considérés pour appliquer

notre modèle qui ne nécessitent pas la prise en compte de tels éléments. Le lien entre les deux ontologies s'effectue donc entre une instance du concept *Objet_commun* de l'ontologie de tâche et un concept de l'ontologie du thème. La modélisation du contexte que nous proposons pourrait cependant être étendue pour prendre en compte le modèle du thème à travers ses instances. Dans ce cas là, le concept « objet commun » aurait pour instances soit des concepts du thème, soit des instances du thème.

2.4.1 Illustration 1 : cas de la veille en astronomie

Dans la figure 4.5, les instances du concept *Objet* de l'ontologie de veille sont trouvées dans l'ontologie du thème de l'astronomie. Le lien établi par *Objet* permet dans ce cas d'établir les concepts de thèmes traités dans un *article* (cf figures 4.1 et 4.3). Les instances sont déterminées à partir des concepts jugés représentatifs des *articles* dans l'ontologie de l'astronomie. Le concept *Objet* permet également d'établir les domaines d'intérêt d'un *chercheur*. Ce lien est établi automatiquement à travers l'analyse des publications dont il est auteur.

L'ontologie de la veille permet quant à elle de stocker, à partir des instances des *articles*, les thématiques abordées et qui sont issues du domaine de l'astronomie. Les thématiques sont mémorisées comme des instances du concept *Objet*.

Dans l'optique d'activités de veille, les liens établis entre les deux ontologies permettent la mise en place de nouvelles analyses [Hernandez 2004c]. Les thématiques de recherche des organismes pourront être analysées à partir des centres d'intérêts des chercheurs appartenant à cet organisme, l'évolution de ces thématiques au cours du temps et l'influence d'un chercheur sur les axes de recherche en considérant la date à laquelle il a rejoint l'organisme.

Des exemples de liens entre les deux ontologies sont présentés dans la figure 4.5. Le *chercheur* de *nom Dupond* et de *prénom Jean a pour domaine d'intérêt* le concept *couronne solaire* du domaine du thème, l'*article* de référence *A-1 traite* du concept *eclipse solaire* du domaine du thème.

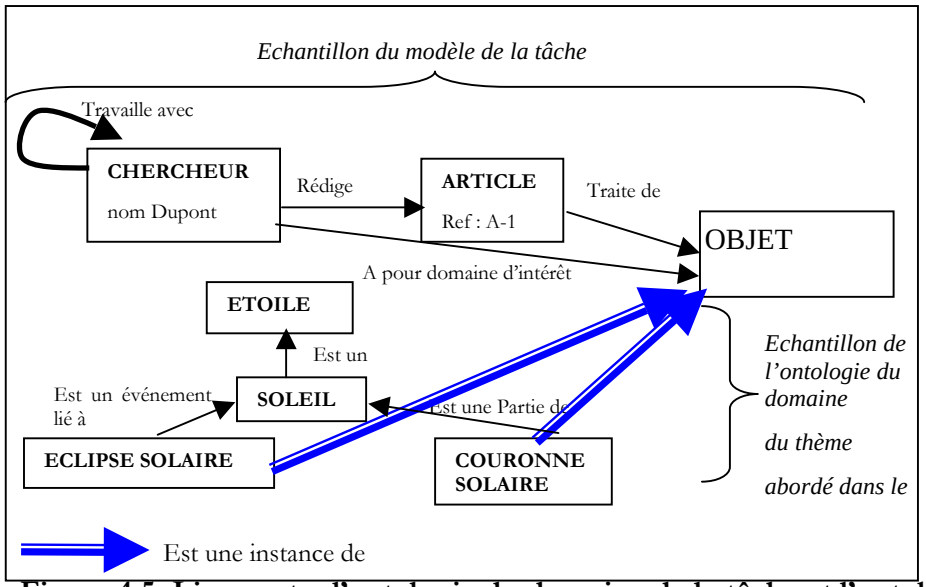


Figure 4.5. Liens entre l'ontologie du domaine de la tâche et l'ontologie du domaine abordé dans le corpus

Les liens entre les deux ontologies se formalisent dans ce cas là par :

$f_{Ctâche} = f_{Ctâche} \cup \{(objet\_commun1, Eclipse\_solaire), (objet\_commun1, Couronne\_solaire)\}$

f_R= f_R∪ { traite_de (article1, Eclipse_solaire), a_pour_domaine (chercheur1, couronne_solaire) }

2.4.2 Illustration 2 : cas de l'apprentissage pédagogique dans le cadre BTS informatique

Dans la figure 4.6, les instances du concept *Objet* de l'ontologie de l'apprentissage pédagogique sont trouvées dans l'ontologie du thème du référentiel du BTS informatique. Le lien établi par *Objet* permet dans ce cas d'établir les concepts de thèmes traités dans un *cours* et un *excercice*.

Des exemples de liens sont proposés dans la figure 4.6. Le *cours* de Référence *C-1* aborde la notion de *composant* du domaine du thème, l'*exercice E-1* qui met en pratique ce cours aborde la notion de *Mémoire*.

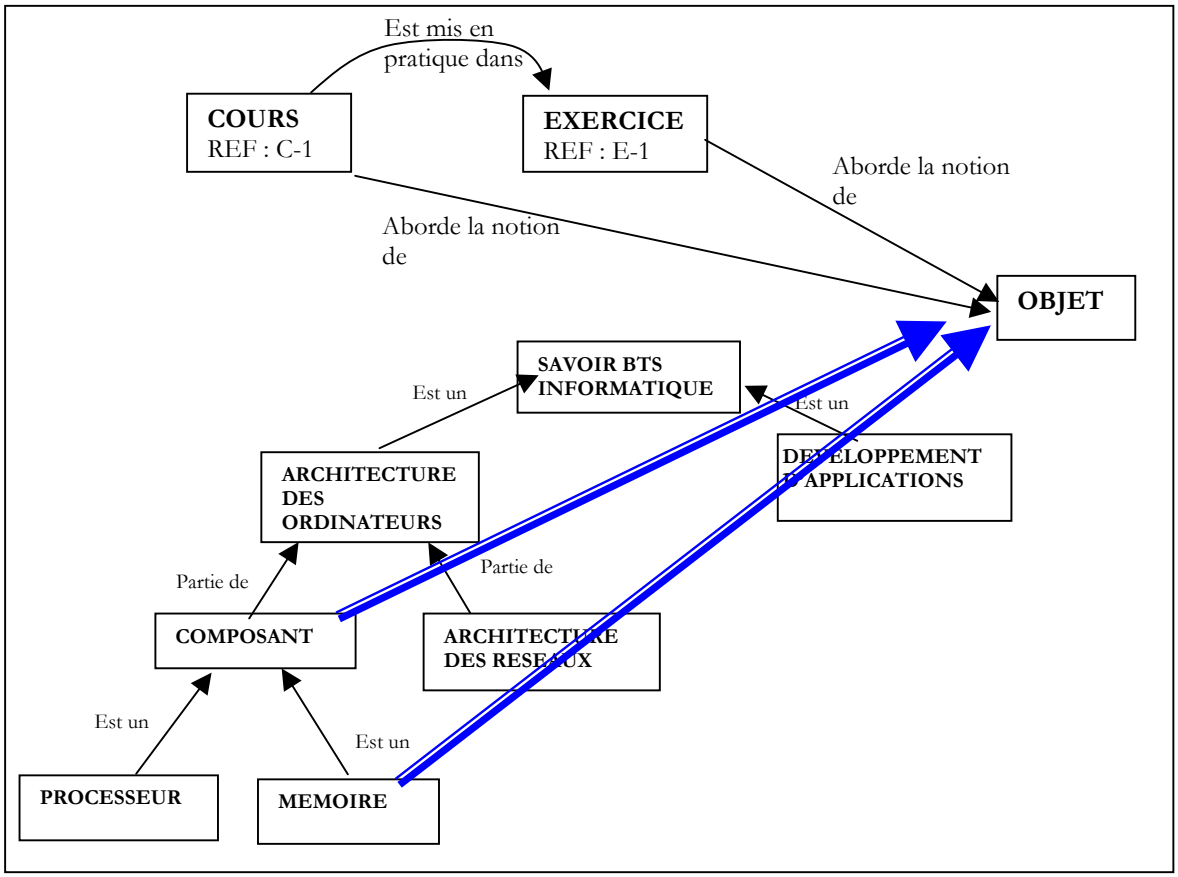


Figure 4.6. Liens entre l'ontologie du domaine de la tâche et l'ontologie du domaine abordé dans le corpus

Le lien entre les deux ontologies se formalise dans ce cas là par :

$f_{Ctâche}$ = $f_{Ctâche}$ ∪ {(objet_commun1, Composant), (objet_commun1, Mémoire)}

f_R= f_R∪ { aborde_la_notion_de (cours1, Composant), aborde_la_notion_de (excercice1, Mémoire) }

3 Intégration du modèle dans un processus de RI

Le contexte représenté à la fois par l'ontologie du thème abordé dans le corpus et l'ontologie de la tâche est intégré dans le processus de RI selon deux perspectives. Dans un premier temps, les granules sont situés par rapport à ce contexte par leur indexation suivant les deux ontologies. Dans un deuxième temps, les deux ontologies spécifiant le contexte sont utilisées comme support à l'accès à l'information. Ces deux aspects seront développés dans les sections 3.1 et 3.2.

3.1 Indexation des granules documentaires

Les deux ontologies représentant le contexte d'une recherche sont utilisées pour indexer les granules. L'indexation joue un rôle primordial dans les SRI et les systèmes d'exploration en définissant les descripteurs qui représentent les granules et à partir desquels les granules peuvent être accédés ou analysés. Deux démarches sont prises en compte pour l'indexation des granules suivant les ontologies du modèle.

La première consiste à indexer sémantiquement le contenu des granules à partir de l'ontologie du thème abordé dans le corpus. Comme nous l'avons présenté dans le chapitre 3, ce type d'indexation s'inscrit dans le prolongement en RI de l'utilisation de ressources terminologiques telles que les hiérarchies de concepts et les thesaurus. Elle vise à choisir les descripteurs du contenu des granules à partir des objets de l'ontologie qu'il référence et non plus à partir des mots qui sont présents [Haav 2001]. L'indexation sémantique n'est possible que par l'existence et l'utilisation de ressources décrivant explicitement l'information correspondant aux objets. L'originalité de l'approche suivie est que cette indexation utilise non seulement les liens hiérarchiques mais également les liens associatifs décrits dans l'ontologie. Ainsi, la dépendance entre les concepts dans l'ontologie et le granule est prise en compte. Les granules pouvant aborder différents concepts, une fonction de pondération des concepts d'indexation est proposée. Elle permet de restituer pour chaque granule les concepts en fonction de leur représentativité du contenu des granules.

La deuxième démarche s'inscrit dans le cadre du Web Sémantique. Dans ce contexte-là, l'indexation (aussi appelée annotation de granules) a un double objectif : indexer le contenu des granules à partir des ressources permettant d'en extraire les concepts et instances mais aussi de représenter les ressources en générant les méta-données correspondantes. Cette démarche repose sur des ontologies modélisant les objets du monde à travers les acteurs et entités que les granules constituent et comportent [Guha 2003]. L'annotation des granules est réalisée, dans notre cas, à partir de l'ontologie du domaine de la tâche et de son lien avec l'ontologie du domaine du corpus.

Ainsi, contrairement aux approches existantes, nous dissocions l'indexation du contenu des granules et celle des objets nécessaires pour réaliser la tâche. Les mécanismes mis en place pour réaliser ces deux types d'indexation sont décrits dans les sections suivantes.

3.1.1 Indexation sémantique des granules à partir de l'ontologie de thème

Afin de déterminer les thématiques abordées dans les documents du corpus, un mécanisme d'indexation sémantique des granules à partir de l'ontologie du thème abordé dans le corpus est mis en place. Ce mécanisme consiste à rechercher les descripteurs du contenu des granules parmi les concepts de l'ontologie. Les granules sont alors indexés non pas à partir de termes qui peuvent être ambigus mais à partir de concepts dont la sémantique est clairement établie. L'indexation sémantique comprend deux étapes : l'identification des concepts dans les granules et la pondération de ces concepts dans le but de déterminer leur représentativité des granules. L'originalité de la méthode proposée est que les concepts sont pondérés par rapport à leur fréquence d'apparition dans le granule et le corpus ainsi que par leur organisation dans la structure de l'ontologie légère. Contrairement aux approches existantes [Desmontils 2004] [Baziz

2005], la pondération prend en compte les relations non taxonomiques de la structure de l'ontologie, ce qui permet un traitement plus approfondi de la sémantique décrite dans l'ontologie.

3.1.1.1 Identification des concepts dans les granules

L'analyseur syntaxique de corpus Syntex présenté dans la section 2.4 du chapitre 2 est utilisé pour extraire l'ensemble des syntagmes de chaque granule. Seuls les syntagmes nominaux sont considérés car ils correspondent à la nature des labels des concepts dans l'ontologie. Syntex extrait les syntagmes d'un granule sous différentes formes : la forme maximale correspond au syntagme composé de toutes les expansions liées au terme recteur du syntagme ; les formes réduites correspondent aux différents syntagmes issus du syntagme maximal pour lesquels les expansions sont successivement supprimées. L'intérêt de considérer les différentes formes sous lesquelles se décompose un syntagme est que les labels des concepts de l'ontologie peuvent être recherchés parmi ces formes. Par exemple, dans la phrase « magnetic connection between black holes and disks are observed », le syntagme maximal « magnetic connection between black holes and disks » est extrait. Un des termes recteurs est "connection" ; il donnera lieu au syntagme réduit "magnetic connection". De la même manière, d'autres syntagmes réduits seront « black hole » et « disk ». Si le syntagme maximal n'apparaît pas dans le lexique de l'ontologie, les syntagmes réduits en feront peut-être partie. Notons qu'afin de capturer les variations lexicales des syntagmes, ceux-ci sont extraits sous leur forme lemmatisée par Syntex. Finalement, le lexique d'un granule est défini par l'ensemble des syntagmes nominaux lemmatisés extraits sous leurs différentes formes.

Ainsi le lexique d'un corpus est noté par :

$$\mathbf{L_{corpus} = \{syntagmes\ lemmatisés\ extraits\}}$$

La phase d'identification des concepts référencés dans un granule consiste à retrouver, parmi les syntagmes composant le lexique des granules, les labels de l'ontologie ainsi que les concepts correspondants. Ces deux étapes sont décrites dans les sections suivantes.

Identification des labels

L'étape d'identification des labels consiste à rechercher, dans le lexique du granule, les syntagmes correspondant à des labels de l'ontologie. Ceci revient à évaluer :

$$\mathbf{L_{Cthème} \cap L_{corpus}}$$

Cette recherche permet d'identifier les labels se rapportant aux concepts les plus spécifiques. En effet, les labels de l'ontologie sont recherchés dans l'ensemble des syntagmes de chaque phrase du granule. Les syntagmes extraits pour chacune des phrases regroupent les syntagmes maximaux et les syntagmes réduits de cette phrase. Afin que les labels les plus spécifiques soient identifiés, les labels retenus sont ceux correspondant aux syntagmes sous leur forme la plus longue (c'est-à-dire comportant le plus de mots).

Par exemple, de la phrase « *magnetic connection between black holes and disks are observed* » appartenant à un granule, les syntagmes : « *magnetic connection between black hole and disk* », « *magnetic connection* » « *connection* », « *black hole* », « *hole* » et « *disk* » sont extraits. Supposons que « *magnetic*

connection », « *connection* », « *black hole* » et « *hole* » soient labels de concepts dans l'ontologie, alors, les labels les plus spécifiques identifiés dans la phrase du granule seront « *black hole* » et « *magnetic connection* ».

Identification des concepts

Cette étape consiste à rechercher pour chaque label le concept associé :

$$c/\ F(l)=c \text{ pour } l \in L_{Cthème} \cap L_{corpus}$$

Dans le cas où le label retrouvé dans un granule permet de référencer un seul concept, le concept correspondant est identifié.

Dans le cas où le label retrouvé dans un granule correspond à plusieurs concepts, un mécanisme de désambiguïsation est mis en place afin de déterminer le concept effectivement référencé dans le granule. Il prend en compte tout d'abord le contexte d'apparition du label dans la phrase à partir des autres concepts non ambigus retrouvés. Si aucun concept non ambigu n'a été retrouvé dans la phrase, le label est analysé par rapport aux concepts retrouvés dans le granule. Le concept choisi est celui qui est sémantiquement le plus proche des autres concepts identifiés dans le contexte documentaire considéré. Afin d'identifier le lien sémantique entre les différents concepts candidats et les concepts du contexte, la distance entre les concepts est évaluée à partir du nombre de relations du plus court chemin séparant les concepts dans l'ontologie [Rada 1989]. Cette mesure est présentée en détail dans la section 1 du chapitre 3. Pour chacun des concepts candidats, la somme du nombre de relations est calculée. Le concept proposé est celui pour lequel cette distance est minimale. Le choix du concept est ensuite validé par l'utilisateur.

Considérons, par exemple, la phrase suivante extraite d'un granule : « ***Polarization*** varies noticeably with emergent *photon* energy below 40keV, being up to 30% and down to -10% for different angles of view; these variations cover the range of observed *magnitudes*.»

Le label « *polarization* » permet de référencer différents concepts : « *polarization concerning wave* », « *polarization concerning charge separation* ». Les labels non ambigus "*photon*" et "*magnitude*" sont extraits de cette phrase, ils référencent les concepts "*photon*" et "*magnititude*". Afin d'identifier le concept référencé par le label ambigu "*polarization*", le nombre de relations séparant les deux concepts qu'il référence (« *polarization concerning wave* », « *polarization concerning charge separation* ») et les autres concepts identifiés dans la phrase (« *photon »* et « *magnitude »*) est calculé. Le nombre de relations séparant « *polarization concerning wave* » et « *photon* » est de 4 car « *polarization concerning wave* » est le père de « *X Ray polarisation* », lui-même lié par la relation « *est un phénomène lié à* » à « *X Ray* », lui-même sous-fils du concept « *photon* ». Le nombre de relations séparant « *polarization concerning wave* » et « *magnitude »* est de *7*. La somme des relations séparant le concept « *polarization concerning wave* » et les concepts identifiés dans le contexte documentaire est donc de 11. Le même calcul est effectué pour le concept « *polarization concerning charge separation »* et la somme est de 30. Le concept choisi pour référencer le label « polarization » dans la phrase est donc le concept « *polarization concerning wave* » car son lien sémantique entre les autres concepts de la phrase est plus fort.

A la fin de cette étape, les concepts référencés dans l'ensemble des granules sont identifiés.

3.1.1.2 Pondération des concepts

La pondération d'un concept vise à déterminer le degré par lequel ce concept de l'ontologie est représentatif d'un granule donné, mais reflète également son pouvoir discriminant c'est-à-dire sa capacité à distinguer les granules pertinents des granules non pertinents lorsque ce concept est considéré. Ce degré est appelé représentativité statistique du concept dans le granule. Les concepts ne sont indépendants dans leur utilisation ni dans les granules ni dans l'ontologie. Le degré de représentativité sémantique prend en compte cette dépendance.

Nous proposons une mesure de pondération des concepts qui prend en compte à la fois la représentativité statistique et la représentativité sémantique du concept dans le granule.

La **représentativité statistique** est calculée par l'adaptation de la mesure tf.idf utilisée en RI pour calculer le pouvoir discriminant d'un terme. Appliquée aux concepts, cette mesure vise à favoriser les concepts apparaissant fréquemment dans le granule mais faiblement dans le reste de la collection. La formule proposée est la suivante :

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

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

où $cf_{c,d}$ représente la fréquence d'apparition des labels du concept c dans le granule g

et f_c correspond au nombre de granules contenant au moins un des labels du concept c

La **représentativité sémantique** prend en compte le lien dans l'ontologie entre le concept considéré et les autres concepts du granule et illustre le contexte sémantique du concept. Elle repose sur le principe suivant lequel plus un concept est proche sémantiquement des autres concepts retrouvés, plus il est représentatif de l'ensemble des thématiques du granule. Cette représentativité sémantique est calculée à partir de la proximité du concept considéré avec les autres concepts retrouvés dans le granule. Elle se calcule par la formule (2).

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

où GC(g) représente l'ensemble des concepts retrouvés dans le granule g.

la mesure de proximité prox est présentée dans la section suivante.

Afin de combiner les deux facteurs de la **représentativité d'un concept**, un concept est pondéré par la somme de sa représentativité statistique et de sa représentativité sémantique, chacune étant normalisée sur l'ensemble de la collection. Les facteurs α et β permettent de leur faire jouer un rôle non symétrique. Cependant, nous n'avons pas étudié l'impact de cette pondération.

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

Il faut noter que ces formules sont génériques car un granule peut correspondre soit à un document, soit à une partie de granule, soit à un regroupement de documents. Ainsi, il est possible de calculer la représentativité d'un concept pour un extrait, ou, au contraire, pour un groupe de documents (éventuellement le corpus complet).

3.1.1.3 Mesure de proximité entre concepts

Afin d'évaluer la proximité sémantique entre concepts, la structure de l'ontologie et les informations contenues dans le corpus sur les concepts doivent être pris en compte. Ainsi, les mesures de la littérature que nous estimons les plus adaptées sont celles qui reposent sur le contenu en information des concepts [Resnik 1995] [Jiang 1997] [Lin 1998]. Ces mesures sont décrites dans le chapitre 3 section 1. Dans ces mesures, l'information portée par les concepts est capturée par la probabilité d'obtenir les concepts dans un corpus. La proximité ou similarité sémantique entre les concepts est ensuite évaluée à partir de leur information commune. Les mesures sont définies pour évaluer la similarité entre concepts dans une taxonomie. L'information commune est reflétée par le contenu en information du concept le plus spécifique qui généralise les deux concepts. Ces mesures s'appliquent au cas de l'évaluation de la proximité entre concepts choisis pour l'indexation de documents dans la mesure où elles évaluent les liens sémantiques à partir de la structure hiérarchique de l'ontologie et de l'information portée par les concepts dans le corpus. Contrairement à la mesure utilisée dans [Desmontils 2002] qui attribue le même poids à toutes les relations « est un », ces mesures permettent également de contrecarrer la subjectivité dans le degré sémantique des liens « est un ».

Cependant, ces mesures ne sont définies que pour la prise en compte des relations « est un ». Or, les relations présentes dans une ontologie légère sont de plusieurs types (taxonomiques, méronymiques, causales, transitives ou sans propriété logique). Ne pas considérer ces relations associatives revient à ignorer une partie de la sémantique contenue dans l'ontologie. Nous proposons donc une extension aux mesures reposant sur le contenu en information des concepts en prenant en considération ces relations.

Les mesures reposant sur le contenu en information des concepts évaluent la similarité entre deux concepts à partir du contenu en information du concept qui les généralise. Ce concept est obtenu en considérant le concept le plus spécifique auquel sont liés les concepts par des relations « est un ». Afin d'étendre les mesures aux relations non-taxonomiques, une première solution suivie dans [Lord 2003] est de considérer les relations associatives comme des relations « père-fils », en quelque sorte en les transformant en relation « est un ». Cependant, cette solution revient à pouvoir considérer un concept généralisant un concept donné sans que ce concept n'ait aucun lien sémantique avec le concept en question. Prenons pour illustrer ce cas de figure, l'exemple présenté dans la figure 4.7 pour lequel les relations associatives (« *appartient à* », « *possède une* ») ont été transformées en relation « père/fils ». Il paraît intuitivement faux de considérer que le concept « Adresse » généralise les concepts « Voiture » et « Magasin ». Le lien sémantique entre les concepts considérés et le concept généralisant est préservé dans les formules initiales par la transitivité de la relation « est un ».

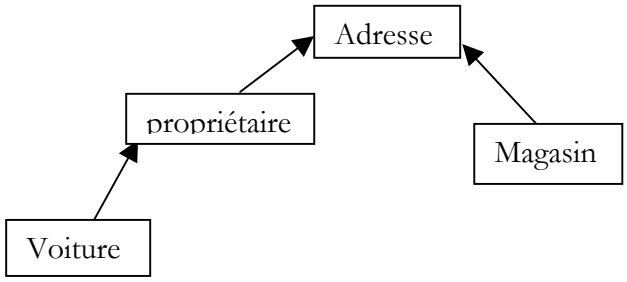


Figure 4.7 Exemple de relations associatives transformées en relation « père/fils »

Notre proposition consiste à adapter la mesure proposée par Jiang [Jiang 1997]. Le choix de cette méthode a été motivé par l'intuition, qui sous-tend cette mesure, selon laquelle les relations ne représentent pas le même poids sémantique en fonction de leur nature (relations taxonomiques, non taxonomiques, …). La distance sémantique représentée par la relation « est un » paraît intuitivement plus importante que celle entre deux concepts liés par un autre type de relation.

La mesure proposée par Jiang vise à quantifier la distance sémantique entre deux concepts à partir du poids des relations sémantiques séparant les concepts par le plus court chemin. Bien que cette mesure permette de différencier le poids des relations, seule la pondération des relations taxonomiques a été proposée.

Les formules (4) et (5) (6) (7) et (8) estiment cette distance.

$$\text{Dist}(c1,c2)=\sum_{c\in pcc(c1,c2)} Poids(c, père(c)) \quad \textbf{(4)}$$

Où pcc(c1,c2) représente l'ensemble des concepts du plus court chemin.

père(c) est le concept parent de c par une relation taxonomique

$$\text{Poids}(c,\text{père}(c))=\text{CI}(c)-\text{CI}(p) \quad \textbf{(5)}$$

$$\text{CI}(c) = -\log (p(c)) \quad \textbf{(6)}$$

$$\text{Freq}(c)=\sum_{n\in word(c)} count(n) \quad \textbf{(7)} \qquad p(c)=\frac{freq(c)}{N} \quad \textbf{(8)}$$

où word(c) représente l'ensemble des termes ou labels représentant les concepts c et les concepts subsumés par c, count(n) le nombre d'occurrences du mot n dans le corpus et N le nombre total d'occurrences de labels de concepts retrouvés dans le corpus.

Le contenu en information (CI) d'un concept est calculé à partir de la probabilité d'apparition dans le corpus des labels du concept ainsi que de ceux de ses concepts descendants (obtenus par la relation « est un ») [Resnik 1995].

Notons que dans notre cas, la fréquence d'un concept est calculée après sa désambiguïsation. Contrairement à la formule proposée par Resnik, le contenu en information d'un concept n'est pas biaisé par la prise en compte d'occurrences de labels pouvant référencer un autre concept.

Jiang propose de simplifier le calcul de la distance entre deux concepts par la formule suivante :

$$\text{Dist}_{simp}(c1,c2)=\text{CI}(c1)+\text{CI}(c2)- 2*\text{CI}(\text{SPS}(c1,c2)) \quad \textbf{(9)}$$

où SPS(c1,c2) représente le concept le plus spécifique subsumant c1 et c2

La méthode de [Jiang 1997] et celle de [Lord 2003] ont chacune leur intérêt. Cependant, Jiang ne considère que les relations taxonomiques et Lord considère toutes les relations mais ne les différencie pas. Nous proposons une nouvelle mesure inspirée de ces deux approches.

Chemin valide

La solution que nous proposons est de considérer le lien sémantique entre deux concepts seulement à partir de chemins jugés valides par le type de leurs caractéristiques (transitivité, symétrie, …) lorsqu'elles sont précisées dans la formalisation de l'ontologie. Nous considérons qu'un chemin est valide s'il n'implique pas plus d'un changement d'orientation. Un changement d'orientation est représenté par une propriété non transitive. Par exemple, en considérant une relation transitive R et les faits suivants : aRb et bRc, on peut inférer que aRc, ceci signifiant que les a et c sont liés par la même propriété. Par contre, dans le cas où la relation est non transitive, ce type d'inférence ne peut avoir lieu et la relation implique un changement d'orientation entre le concept a et le concept c. La prise en compte du changement d'orientation est importante dans le cadre des mesures reposant sur le contenu en information des concepts car elle permet de garantir que les concepts considérés ont des caractéristiques communes. Dans le cas où le lien est symétrique, la relation est considérée dans les deux sens. Le chemin impliquant les deux concepts est jugé valide dans les deux sens.

Mesure

Dans le cas des relations « est un », le contenu en information des concepts est calculé à partir de leurs descendants issus des relations « est un », comme proposé à l'origine dans [Resnik 1995]. Dans le cas des relations non taxonomiques, le poids du lien ne peut être évalué par la formule (4) proposée par Jiang car les contenus en information des deux concepts n'ont aucun élément commun. Le contenu en information n'est, en effet, capturé à partir d'aucun label commun. Nous proposons donc d'évaluer le poids de ces relations à partir d'un nouvel élément appelé contenu en information de la relation. Le contenu en information de la relation est calculé à partir de la probabilité de co-occurrence des deux concepts dans le corpus. Cette probabilité est établie à partir du nombre de fois où les labels du concept c1 ainsi que les labels de ces descendants co-occurrent avec des labels du concept c2 et de ses descendants. De la même façon que pour le contenu en information des concepts, la fonction –log est utilisée pour réduire le contenu en information d'un lien ayant une forte probabilité. Les formules (10) et (11) permettent de calculer le poids d'un lien non taxonomique.

$$Poids_{non\_taxo}(c,c_2)=CI(c,c_2) \quad \textbf{(10)}$$

$$\text{Avec } CI(c,c2)=-\log\left(\frac{nbdoc\_cooc(c,c2)}{nbdoc\_total}\right) \quad \textbf{(11)}$$

où nbdoc_cooc(c,c2) est le nombre de documents dans lesquels c et c2 ou chacun de leurs descendants co-occurrent et nbdoc_total est le nombre total de documents du corpus

La formule proposée pour calculer la distance entre deux concepts c1 et c2 est alors la suivante :

$$dist_{prop}(c1,c2)=\sum_{c\in pccvalide(c1,c2)} Poids_{prop}(c,rel(c)) \quad \textbf{(12)}$$

où pccvalide(c1 et c2) est le plus court chemin valide entre c1 et c2

et rel(c) est le concept lié à c dans l'ontologie.

$$Poids_{prop}(c,c_2)=CI(c)-CI(c_2) \text{ si } c_2=père(c) \quad \textbf{(13)}$$

$$Poids_{prop}(c,c_2)=CI(c,c_2) \text{ sinon}$$

La proximité est ensuite calculée à partir de l'inverse de la distance à partir de la formule (14).

$$\text{Prox}_{2\_\text{prop}}(c1,c2) = \frac{1}{1+dist_{prop}(c1,c2)} \quad \textbf{(14)}$$

Une évaluation de la mesure est proposée dans le chapitre 7.

3.1.2 Annotation des documents à partir de l'ontologie de tâche

L'annotation des granules suivant l'ontologie du domaine de tâche consiste à rechercher dans les granules les valeurs des méta-données représentées dans l'ontologie de tâche. Les valeurs des méta-données peuvent être explicitement présentes dans les granules, les techniques mises en place sont alors celles présentées dans la section 3.1.2.1. Les valeurs peuvent être extraites à partir de l'ontologie du domaine traité dans le corpus, les techniques employées sont décrites dans la section 3.1.2.2.

3.1.2.1 Extraction des instances présentes dans les granules

Des techniques d'extraction de connaissances sont employées pour extraire les instances de concepts correspondant à des méta-données explicitement présentes dans les granules. Dans notre cas, un mécanisme d'extraction est mis en place à partir de l'analyse des balises des granules. Une correspondance manuelle est établie entre les balises et les concepts correspondant aux méta-données explicitées soit à partir de la DTD si les documents sont représentés en XML soit à partir du format de représentation propre au corpus. Pour chacune des chaînes de caractères encadrées par les balises, une instance du concept correspondant est créée. Dans le cas où les granules ne seraient pas structurés explicitement, des techniques plus sophistiquées d'extraction d'information pourraient être utilisées [Dkaki 1997] [Amous 2001]. D'autre part, dans le cas des granules issus d'un document (partie de document), les méta-données sont implicitement extraites des méta-données associées au document. Dans le cas des granules composés de plusieurs documents, un mécanisme d'inférence des méta-données est nécessaire.

3.1.2.2 Extraction des instances d'un objet commun aux deux ontologies

Nous nous plaçons ici dans le cas où les instances du concept objet de l'ontologie de tâche sont implicitement présentes dans les granules. Elles sont extraites à partir de l'analyse des granules contenant l'information ciblée par l'objet.

Cet objet prenant ses valeurs dans l'ontologie de thème, elles sont déterminées par l'analyse de la représentativité des concepts de l'ontologie du thème dans l'ensemble des granules considérés. L'indexation sémantique utilisée est celle présentée dans la partie 3.1.1.

Par exemple, dans le cas de l'ontologie du domaine de la tâche de la veille présentée dans la figure 4.1, ces instances correspondent à celles du concept *« objet »*. Dans le cas où les instances recherchées sont celles relatives au domaine d'intérêt d'un chercheur, les instances correspondant au concept *« objet »* sont les concepts jugés représentatifs de ses publications. Ces concepts sont extraits en fonction de leur représentativité. L'ensemble des granules considérés n'est plus le corpus mais l'ensemble des publications du chercheur.

L'originalité de l'approche est que les concepts extraits de l'ontologie de thème sont pondérés en fonction de leur représentativité. Dans le cas où un granule ou un ensemble de granules aborde de nombreux concepts, les concepts sont organisés en fonction de ce score. Ces concepts pouvant être des instances de l'ontologie de tâche, un poids est donc aussi associé à ces

instances. Ceci permet à un système d'accès à l'information de restituer les instances en fonction de leur pertinence en ciblant l'information restituée.

3.2 Accès à l'information

L'intérêt de notre approche est que le contexte est pris en compte pour fournir à l'utilisateur un accès à l'information contenue dans le corpus [Hernandez 2005b]. Dans notre approche, le contexte d'une recherche correspond à l'ontologie du thème et à l'ontologie de la tâche. L'utilisateur a accès à l'information à partir d'une présentation de l'information par la navigation dans les deux ontologies. L'utilisateur est alors guidé à partir d'éléments cibles de sa recherche. L'accès à l'information peut être réalisé à deux niveaux d'abstraction : le niveau des concepts et le niveau des instances. D'autre part, thème et tâche peuvent être combinés pour une meilleure formulation du besoin en information.

3.2.1 Concepts / instances

Un premier type d'accès à l'information propose deux niveaux d'abstraction.

Le premier niveau est réalisé à partir des concepts des deux ontologies. Ainsi, par leur visualisation, l'utilisateur peut avoir une vue d'ensemble sur la collection et la connaissance qui lui est associée. En particulier, sa connaissance a priori du thème pourra être confrontée à celle présente dans l'ontologie associée. Cette présentation permet de l'affranchir de certaines erreurs d'interprétation sémantique et peut l'aider à mieux déterminer et formuler son besoin.

Le second niveau s'effectue à partir des instances de l'ontologie de tâche. Ceci permet de rendre accessible à l'utilisateur les détails sur le contenu de la collection. Il peut ainsi visualiser les objets de la collection et leurs liens.

Les deux niveaux sont accessibles l'un par l'autre ; ainsi, le choix d'un concept peut amener à la visualisation de ses instances. Réciproquement, à partir d'une instance, il est possible de connaître le rôle du concept associé dans la tâche.

3.2.2 Thème/tâche

Un deuxième accès consiste à présenter les informations du corpus suivant les deux types d'information qui permettent à l'utilisateur de spécifier son besoin. D'une part, il peut explorer la collection à partir des concepts de l'ontologie du thème. Il peut ainsi spécifier son besoin en fonction des thématiques abordées dans la collection. Il peut naviguer dans les concepts, en choisissant des concepts plus génériques/plus spécifiques, ou associés et ne perd jamais ainsi le contexte de sa recherche. D'autre part, il peut accéder directement aux informations qui l'intéressent en explorant l'ontologie du domaine de la tâche. Les concepts représentent l'objet général de la recherche, alors que les instances représentent leur explicitation dans la collection. Cet accès lui permet d'avoir une vue globale sur la collection (valeurs manquantes, ou au contraire grand ensemble de valeurs), mais également de visualiser une donnée spécifique à travers les instances d'un concept. Les informations du corpus étant établies à partir des deux ontologies, les liens entre elles lui sont présentés de façon interactive.

Il est difficile d'expliquer ces mécanismes en dehors d'un exemple concret. Le prototype d'interface reposant sur les mécanismes définis à partir des ontologies que nous avons réalisé dans le cadre de cette thèse est présenté dans le chapitre 8.

4 Conclusion

Le modèle du contexte que nous avons présenté dans ce chapitre se focalise sur deux aspects du contexte : le thème de la recherche et la tâche. Nous avons choisi une modélisation reposant sur des ontologies légères. La formalisation permet une indexation sémantique plus élaborée. Cet enrichissement de l'indexation est renforcé par notre choix d'utiliser des ontologies de domaine plutôt que des ontologies génériques telles que WordNet. Les ontologies de domaine limitent également les problèmes liés à l'ambiguïté des termes. Par ailleurs, contrairement aux rares approches qui s'intéressent à la tâche, nous avons choisi de distinguer la modélisation de la tâche et du thème au travers de deux ontologies. Associée à la formalisation, cette distinction les rend réutilisables pour différentes tâches et différents domaines. Les ontologies sont de plus articulées l'une par rapport à l'autre par l'instantiation d'un concept de l'ontologie de tâche dans l'ontologie de thème. Le rôle des concepts du thème est alors précisé à partir du contexte de la tâche de recherche effectuée. Le mécanisme de représentation des documents à partir des ontologies que nous avons mis en place s'inscrit dans le domaine du Web Sémantique par la proposition d'une mesure visant à pondérer le lien entre un concept de l'ontologie de domaine et les granules documentaires. Cette pondération a l'originalité de prendre en compte à la fois les fréquences d'apparition des concepts et leur organisation dans l'ontologie légère. Notre modèle enrichit d'autre part les types d'accès traditionnels en permettant à la fois une visualisation globale du contenu de la collection et un accès spécifique aux données intéressant l'utilisateur. Le prototype présenté dans le chapitre 8 implante l'intégration du modèle proposé dans un SRI.

La modélisation du contexte proposée permet d'entrevoir de nombreuses extensions. Des agents d'analyses intelligents pourraient être intégrés au modèle de représentation de la tâche. Par exemple, les mécanismes d'analyse tels que ceux décrits dans [Dousset 2003] (classification de documents, analyse de données géoréférencées, etc …) pourraient être intégrés dans la tâche de veille. La création de profils utilisateur est aussi une perspective envisagée. Ces profils permettraient de personnaliser la navigation en fonction de caractéristiques connues sur un utilisateur ou choisies par lui (niveau de profondeur des concepts visualisés, choix des relations à afficher, concepts connus, documents déjà explorés …).

Une limite de notre approche concerne l'accès aux corpus multi-domaines. Il est cependant envisageable de faire co-exister plusieurs ontologies de thème pour répondre à ce type d'environnement.

Nos contributions décrites dans ce chapitre ont été diffusées et reconnues par la communauté scientifique à travers différentes présentations et publications nationales et internationales [Hernandez 2003b], [Hernandez 2004b], [Hernandez 2004c], [Hernandez 2005a], [Hernandez 2005b].

Comme nous l'avons indiqué en introduction de ce chapitre, les ontologies que nous utilisons peuvent pré-exister. Elles peuvent aussi être construites à partir d'un thésaurus. Nous proposons dans le chapitre 5 une méthode pour permettre cette construction. Dans tous les cas, l'ontologie de thème doit être adaptée au corpus ; le chapitre 6 présente une méthode de mesure d'adéquation d'une ontologie à un corpus.

Chapitre 5

D'un thesaurus vers une ontologie légère de domaine une méthode

1 Introduction

De nombreux thésaurus ont été créés dans différents domaines dans l'objectif de proposer un vocabulaire contrôlé pour l'indexation de ressources documentaires et pour l'aide à la formulation d'une requête par un documentaliste. Ils ont nécessité de lourds efforts pour leur conception manuelle. L'existence de normes (ISO 2788 et ANSI Z39) permet d'uniformiser leur contenu en termes de liens sémantiques entre unités lexicales (synonymie, liens hiérarchiques et d'association). Cependant, leur format n'est pas normalisé : fichiers ascii, html, bases de données co-existent. Pour faire face à ce problème, les normes en cours d'élaboration dans le cadre du W3C comme SKOS Core[1] visent à faire migrer les thésaurus vers des ressources disponibles sur le Web Sémantique en se basant sur le langage OWL. La disponibilité de telles ressources sous format normalisé est un enjeu important dans le domaine de la RI. Comme nous l'avons précisé dans le chapitre 2, elle présente principalement trois avantages. Le premier est que l'uniformisation de leur représentation à partir de langages dédiés au Web Sémantique (tel que RDF et OWL) permettra à ces ressources d'être distribuées sur le Web. De plus, ces ressources pourront être uniformément manipulées à partir d'outils dédiés aux ontologies pour leur visualisation, l'annotation, etc… Enfin, les processus de RI pourront s'appuyer sur ces ressources élémentaires simples, sans avoir à faire face à l'hétérogénéité des formats. D'un point de vue de la représentation des connaissances, les thésaurus ont un faible degré de formalisation. Ce sont des collections de termes qui sont organisées suivant une ou plusieurs hiérarchies avec des relations entre termes. Les thésaurus n'ont pas de niveau d'abstraction conceptuelle [Soergel 2004]. 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 » [Fischer 1998]. La relation associative «est lié à» est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques [Tudhope 2001]. Par exemple, dans le thésaurus BIT[2] relatif au monde du travail, le terme « famille » est lié aux termes « femme » et « congé familial », la relation sémantique entre ces deux paires de termes est intuitivement différente. Par les choix faits lors de leur conception, les thésaurus manquent de formalisation et de cohérence par rapport aux ontologies légères.

Les ontologies légères ou lourdes ne posent pas ce type de problème. Elles sont supposées respecter la relation de subsomption dans l'organisation hiérarchique des concepts. D'autre part, les liens d'associations entre concepts sont sémantiquement mieux décrits. Cependant, leur élaboration est coûteuse : elle nécessite de nombreuses interventions manuelles. En effet, les techniques de construction d'ontologies de la littérature (cf section 2 du chapitre 2) reposent généralement sur aucune connaissance préalable du domaine.

Notre approche vise au contraire à réutiliser les thésaurus de domaine qui ont nécessité de lourds efforts de conception pour l'élaboration de nouvelles ressources d'un niveau formel plus élevé. La conception d'ontologies à partir d'un thésaurus présente l'avantage de reposer sur l'ensemble des termes qu'il contient et qui ont été identifiés par des experts comme étant

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

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

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 dans ce chapitre une méthode pour transformer un thésaurus en ontologie légère de domaine pour la RI en plusieurs étapes. Cette méthode vise à s'appliquer à n'importe quel thésaurus de domaine conçu sous les normes ISO 2788 et ANSI Z39. Ces thésaurus sont monolingues et ne sont pas organisés suivant des facettes.

Les approches de la littérature se contentent d'étudier les mécanismes de transformation, sans s'intéresser à la mise à jour de la connaissance en phase d'utilisation du thésaurus. Cette mise à jour correspond à une autre contribution de cette thèse. La méthode que nous proposons vise d'une part à expliciter la connaissance représentée dans le thésaurus existant et d'autre part à la mettre à jour à partir d'informations extraites de textes.

Nous illustrons nos propositions à partir du thésaurus de l'astronomie IAU[3]. Ce cas d'étude est également utilisé dans le cadre de leur évaluation dont les résultats sont présentés dans le chapitre 7.

La section 2 présente la méthode que nous proposons. Cette section explicite les problématiques auxquelles la méthode doit répondre, les différentes étapes qu'elle met en place, ainsi que le schéma conceptuel de l'ontologie choisi. Les sections suivantes décrivent les étapes de la transformation de l'ontologie. La section 4 présente les mécanismes utilisés pour créer le niveau d'abstraction conceptuel à partir du thésaurus. La section 5 explique comment la structure de l'ontologie est construite (liens entre concepts). La dernière section expose les techniques employées pour mettre à jour la connaissance représentée dans l'ontologie ainsi construite.

2 Présentation de la méthode

La méthode 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 de la connaissance actuelle d'un domaine, la méthode repose sur l'analyse de documents textuels. Les problématiques auxquelles doit répondre la méthode sont situées dans le cadre général de la construction d'ontologies à partir de textes. Ce cadre général est décrit à partir de la méthodologie TERMINAE dans la section 2.1. La méthode proposée repose sur différentes étapes de transformation précisées dans la section 2.2. Les résultats obtenus à partir des différentes étapes sont représentés dans l'ontologie à partir d'un schéma conceptuel présenté dans la section 2.3.

2.1 Cadre général

La méthode que nous proposons s'appuie sur des documents textuels. La méthodologie TERMINAE que nous avons présentée dans la section 2.1.2 du chapitre 1 décrit les différentes étapes dans la construction d'une ontologie à partir de textes. Nous nous basons sur ces étapes pour spécifier la méthode permettant la transformation d'un thésaurus en ontologie. Afin d'identifier les éléments clés dans la transformation d'un thésaurus, nous reprenons les étapes de la méthodologie et les choix que nous faisons pour chacune d'entre elles.

La première étape 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, les besoins que nous identifions sont les suivants :

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

- la spécification des termes du domaine et de leurs variantes lexicales afin de les détecter dans les granules documentaires,
- le regroupement de ces termes en concepts afin de déterminer les objets et notions référencés dans les documents,
- la structuration des concepts à partir de relations taxonomiques et associatives afin de permettre une indexation sémantique de qualité,
- la formalisation de l'ontologie dans un langage interprétable par le SRI afin qu'il soit capable de la manipuler.

Une ontologie légère créée pour la RI doit donc intégrer ces différents éléments.

La deuxième étape repose sur le choix du corpus de référence à partir duquel l'ontologie est construite. Ce choix est un paramètre déterminant de l'élaboration de l'ontologie [Condamines 2005]. Le corpus doit décrire les éléments de connaissance qui seront intégrés dans l'ontologie. Dans le cas de la transformation d'un thésaurus, le corpus doit répondre à deux conditions. Il doit tout d'abord permettre de capturer la connaissance implicite qui n'est pas formalisée dans le thésaurus. Ensuite, le corpus doit aider à la mise à jour de la connaissance à partir de documents récents du domaine. L'ontologie étant créée pour des activités de RI, le corpus considéré doit aider à préciser le contexte associé à des documents du domaine d'intérêt considéré. Dans notre approche, le corpus est extrait de corpus existants et des experts doivent s'assurer qu'il couvre l'ensemble du domaine sur une période représentative. Des résumés d'articles publiés dans des revues du domaine permettent de décrire ce type d'information. Les articles complets pourraient être utilisés mais l'avantage des résumés est que les informations qu'ils détiennent sont synthétisées.

La troisième étape est celle de l'étude linguistique du corpus. Cette étape vise à extraire 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 non présents dans le thésaurus et les relations entre termes qui n'y sont pas explicitées. Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique Syntex que nous avons présenté dans la section 2.4.1 du chapitre 1. 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). Une méthode doit cependant être élaborée pour définir les mécanismes permettant de sélectionner les termes et leurs relations, à partir de la connaissance extraite du thésaurus et des informations extraites du corpus. La méthode que nous proposons vise à répondre à cette problématique.

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

La dernière étape est celle de la formalisation : le réseau sémantique défini à l'étape précédente est traduit dans un langage formel. La formalisation de l'ontologie créée à partir d'un thésaurus peut être réalisée à partir du langage OWL présenté dans la section 4.3 du chapitre 1. 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.

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

2.2 Etapes de la méthode

La méthode proposée repose sur trois étapes. Ces étapes sont décrites dans la figure 5.1.

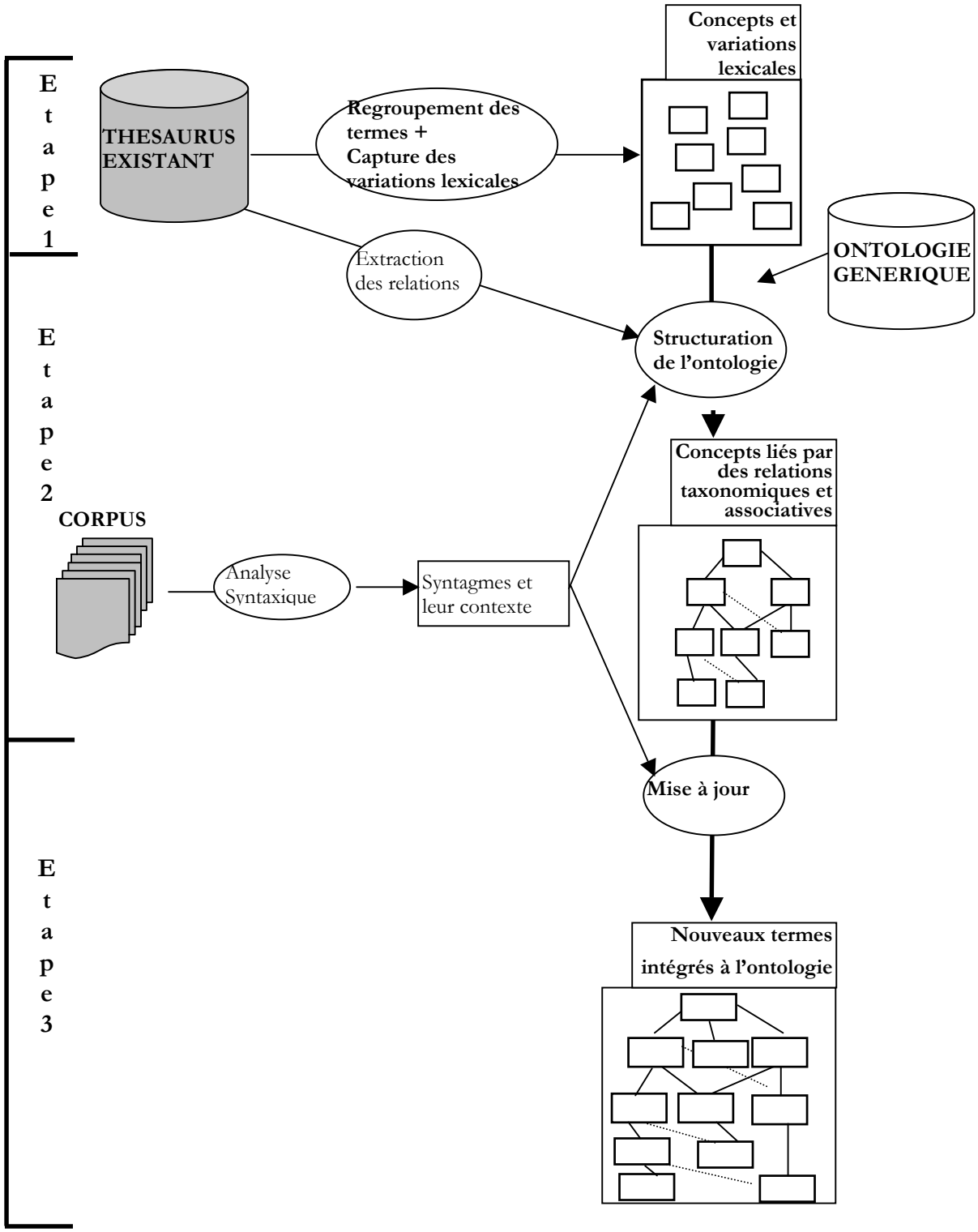


Figure 5.1 Etapes de la méthode

La première étape vise à extraire du thésaurus un ensemble de concepts ainsi que leurs variations lexicales. Peu de méthodes dans la littérature mettent en œuvre cette étape. La majorité

d'entre elles considèrent en effet qu'un concept est référencé à partir d'un seul terme [Sanderson 1999] [Morin 1999] [Maedche 2000] [Ok Koo 2003]. Par l'utilisation d'un thésaurus, notre méthode vise à proposer un mécanisme automatique de regroupement des labels d'un même concept. Cette étape est décrite dans la section 3.

La deuxième étape permet de structurer les concepts de l'ontologie à partir de la détection de relations taxonomiques et associatives dans le thésaurus et dans le corpus. Cette étape soulève différentes problématiques de la construction d'ontologies. L'une d'elles relève de la difficulté à organiser les concepts par des relations taxonomiques. Dans notre cas, les relations hiérarchiques entre termes du thésaurus peuvent être utilisées pour aider à la détection de ces relations. Cependant, un des inconvénients des thésaurus est que le niveau hiérarchique le plus général est souvent composé de nombreux termes. Afin d'organiser les concepts à partir d'un niveau d'abstraction comportant un nombre limité de concepts, nous proposons l'utilisation d'une ontologie générique. Cette ontologie est utilisée pour définir semi-automatiquement les types abstraits du domaine et structurer l'ontologie. Le mécanisme développé est décrit dans la section 4.1. Une autre problématique que fait intervenir cette étape est la détection de relations associatives entre concepts et la désignation de ces relations sémantiques. Peu de méthodes désignent correctement les labels des relations. Nous proposons un mécanisme visant à proposer semi-automatiquement ces relations ainsi que leur label. Le mécanisme repose sur l'analyse syntaxique du corpus de référence qui permet d'extraire les syntagmes constituant le lexique du corpus ainsi que le contexte dans lequel ils apparaissent (noms et verbes qu'ils régissent et par qui ils sont régis). Il est décrit dans la section 4.2.

Finalement, la troisième étape met à jour l'ontologie à partir de la connaissance présente dans les documents et qui n'a pas été extraite du thésaurus. Les problématiques auxquelles répond cette étape sont la détection de termes du corpus de référence à ajouter à l'ontologie et leur intégration semi-automatique dans l'ontologie. Les mécanismes utilisés sont décrits dans la section 5.

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

2.3 Schéma conceptuel

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

Le haut niveau conceptuel est présenté dans la figure 5.2 et est décrit dans les différents sous-paragraphes suivants.

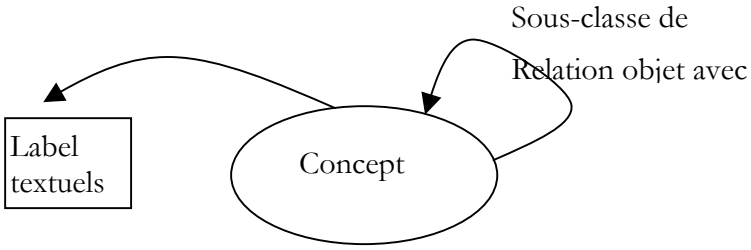


Figure 5.2 Haut niveau du schéma conceptuel de l'ontologie

2.3.1 Concept et label textuel

Un concept est représenté à partir d'une classe OWL `<owl:Class rdf:about="identifiant_unique">`. Celle-ci est identifiée à partir d'un identifiant unique. Les labels d'une classe sont représentés par la propriété label `<rdfs:label>`. Dans le schéma conceptuel proposé dans [Miller 2005], les labels sont de deux types : les labels représentant les termes principaux et ceux représentant les variations lexicales de ces termes. Cette approche peut être intéressante dans le cas où l'application et l'utilisateur doivent avoir une vision différente du contenu de l'ontologie. Dans la mesure où cette différenciation n'a pas lieu d'être dans la recherche de l'adéquation entre une ontologie et un corpus, ni dans le processus de RI, nous avons choisi de ne pas différencier les labels par rapport à leur rôle dans la désignation du concept. Les différentes variations lexicales des termes désignant le concept sont ainsi représentées par cette même propriété.

2.3.2 Relation entre concepts

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

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

2.3.3 Schéma conceptuel d'un thésaurus

L'ensemble des éléments du schéma conceptuel précédemment décrit n'est pas présent dans un thésaurus. Un thésaurus est un ensemble de termes organisé suivant un nombre restreint de relations [Foskett 1980]. Les relations présentes dans un thésaurus répondant aux normes ANSI Z39 et ISO 2788 sont rappelées dans la figure 5.3.

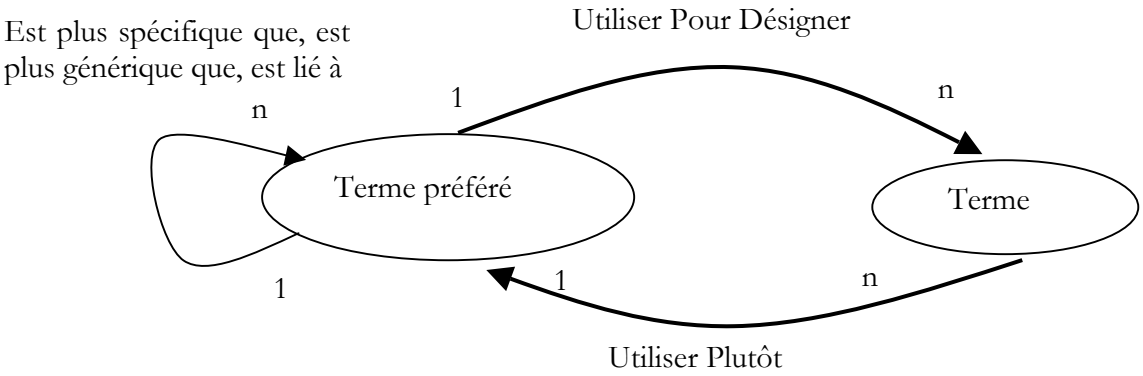


Figure 5.3 Rappel des relations entre termes dans un thésaurus

Dans notre méthode, nous considérons que les thésaurus réutilisés pour construire une ontologie sont de ce type.

Nous faisons trois hypothèses sur la réutilisation des relations entre termes :

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

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

3 Conceptualisation du lexique du thésaurus

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

3.1 Regroupement des termes en concepts

3.1.1 Regroupement reposant sur les relations explicites UP et UPD

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

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

Si t3 UP t1 alors t1 et t3 sont regroupés, avec t1 terme préféré

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

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

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

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

Si t1 UPD t2 et t2 UPD t3, alors t1 UPD t3 => t1, t2 et t3 sont regroupés, avec t1 terme préféré principal

Si t4 UP t5 et t5 UP t6 alors t4 UP t6 => t4, t5 et t6 sont regroupés, avec t6 terme préféré principal

(R2)

La figure 5.4 schématise plusieurs exemples de groupements. Pour faciliter la lisibilité, les termes préférés sont en gras majuscules. Les termes regroupés par R1 sont soulignés en pointillé. Les termes regroupés par la règle R2 sont soulignés en trait plein.

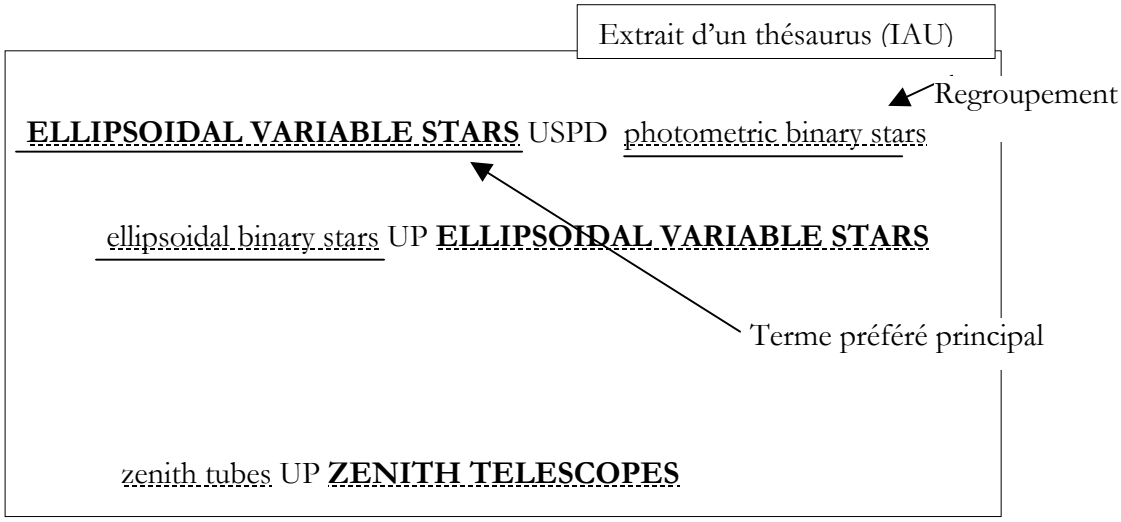


Figure 5.4 Exemples de groupements des termes du thésaurus

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

3.1.3 Identifiant du concept

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

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

3.2 Capture des variations lexicales

La forme lexicale sous laquelle se trouvent les termes du thésaurus est un sujet délicat et largement détaillé dans l'ensemble des normes dédiées aux 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. Si le terme représente une catégorie, le pluriel du terme est préféré, si le terme définit le sens du terme, le singulier est choisi. Les normes ISO 2788 et ANSI Z39 proposent, pour différencier ces cas de figure, la distinction des termes à partir de leur type : les termes désignant des objets dénombrables et les termes désignant des objets indénombrables. Lorsque peut être posée la question « combien d'objets représentés par le terme existent ?», le terme est intégré dans le thésaurus au pluriel, dans le cas contraire il l'est au singulier. Ces règles sont scrupuleusement respectées dans la plupart des thésaurus, comme dans le thésaurus de l'astronomie IAU. Il est cependant possible de trouver des variantes de l'application de ces règles. Par exemple, dans les thésaurus BIT[4] (Terminologie du travail, de l'emploi et de la formation), British Museum[5] et Alcohol and Other Drug Thesaurus[6], les termes sont au singulier sauf si l'usage impose le pluriel.

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

Des techniques de lemmatisation où le recours à un expert peuvent être envisagés. Alternativement, une ressource lexicale telle que WordNet peut être utilisée. La figure 5.5 illustre les concepts identifiés dans la figure 5.4 pour lesquels les labels sont mis au singulier grâce à WordNet.

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

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

Figure 5.5 Exemples de concepts labellisés par des termes au singulier

[4] http://www.ilo.org/public/french/support/lib/indexati/unit1/unit1.htm

[5] http://www.mda.org.uk/bmobj/Objintro.htm

[6] http://etoh.niaaa.nih.gov/AODVol1/titlepage.htm

4 Construction de la structure de l'ontologie

La structure de l'ontologie définit les relations entre concepts établis suite aux étapes présentées dans la section précédente. La structure comprend des relations taxonomiques de type « est un » qui sont décelées par des techniques décrites dans la section 5.1 et des relations associatives qui sont obtenues par les méthodes présentées dans la section 5.2.

4.1 Construction de la hiérarchie de concepts

Certains liens hiérarchiques entre concepts sont directement issus des liens explicites présents dans le thésaurus. Des niveaux hiérarchiques supérieurs y sont ajoutés à partir de l'analyse des têtes et expansions des labels des concepts ainsi que de la création de types abstraits. La figure 5.6 schématise ces différents mécanismes.

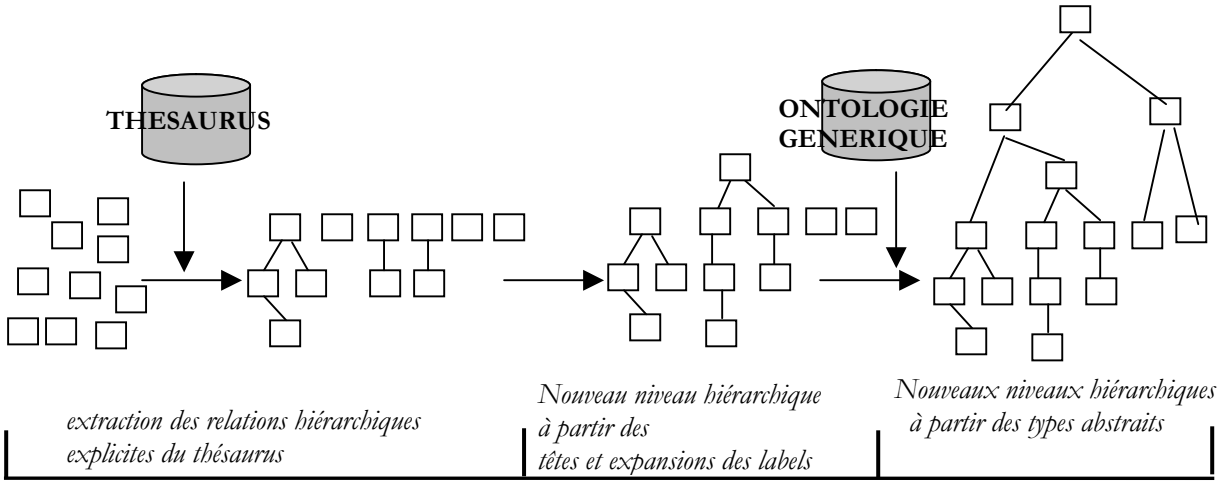


Figure 5.6 Mécanisme de construction de la hiérarchie de concepts

4.1.1 Extraction des relations hiérarchiques explicitées dans le thésaurus

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

Si t1 est plus spécifique que t2 avec t1 label du concept c1 et t2 label du concept c2

=> c1 « est une sous-classe de » c2

(R4)

4.1.2 Suppression de la redondance dans les relations hiérarchiques

Les thésaurus n'étant pas formalisés, des redondances dans la structure hiérarchique de l'ontologie construite avec les règles de R1 à R4 peuvent exister. La relation de généricité est une relation transitive et permet le type d'inférence suivant : si A « est une sous-classe de » B et B « est une sous-classe de » C, alors A « est une sous-classe de » C ; A,B,C étant des concepts. La figure 5.7 en présente un exemple : les flèches entre les rectangles représentant les concepts symbolisant la relation « est une sous-classe de ». Par la propriété de transitivité de la relation « est une sous-classe de », la relation « est une sous-classe de » entre *planetary nebula* et *nebula* est donc inutile.

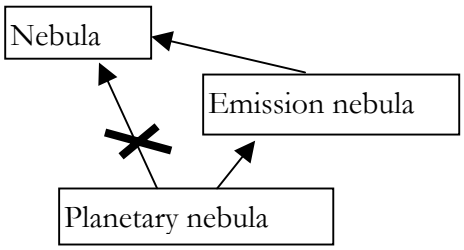


Figure 5.7 Exemple de redondance dans la hiérarchie de l'ontologie

Afin de supprimer les relations redondantes, la pertinence de chacune des relations «est une sous-classe de » est vérifiée. Cette analyse est réalisée grâce à la théorie des graphes. Le graphe G(C,R) où C représente les concepts de l'ontologie et R les relations « est une sous-classe de » est considéré. Le prédicat concepts_accessibles(G,c,R) identifie l'ensemble des concepts qui peuvent être atteints par la fermeture transitive du graphe G à partir du concept c par les relations de R. Pour chacun des concepts, les concepts accessibles CA_{pci} par chacune de ses classes pères p_{ci} sont identifiés par (concepts_accessibles(G,p_{ci},R) $\cap$ p_{ci}). Lorsque l'intersection des concepts accessibles par deux des classes mères d'un concept est non vide ($CA_{pci} \cap CA_{pcj} \neq 0$) , une ou plusieurs relations de redondance sont décelées (le nombre de redondances dépend du nombre de concepts appartenant à $CA_{pci} \cap CA_{pcj}$). Ceci signifie en effet que deux chemins dans le graphe mènent au concept $cinter_k$ appartenant à $CA_{pci} \cap CA_{pc}$. Afin de déterminer les relations à supprimer, le nombre d'arcs des deux chemins séparant le concept c et le concept $cinter_k$ est calculé. Seul le plus long chemin est conservé. Les relations « est un » liant le concept $cinter_k$ à sa classe fille contenu dans le chemin le plus court est alors supprimé. Par l'analyse des classes mères de l'ensemble des concepts de l'ontologie, les relations redondantes sont supprimées.

La suppression de la redondance est formalisée par la règle R5.

Pour tout concept $c \in C$,

Si $\forall\ c_i \in C\ c \neq c_i$, $\exists$ chem1, chem2 tel que chem1=chemin(c,c_i) et chem2=chemin(c,c_i), avec chem1 $\neq$chem2

=> suppression de l'arc à l'origine du chemin le plus court

(R5)

4.1.3 Nouveaux niveaux hiérarchiques

Une des lacunes des thésaurus est que leur plus haut niveau hiérarchique contient

généralement un très grand nombre de termes [Soergel 2004]. Ces termes sont ceux pour lesquels aucune relation « est plus spécifique » n'a été définie. Ceci s'explique par le fait que les thésaurus ne définissent pas de catégories génériques permettant de répertorier l'ensemble des termes du domaine. Cette même lacune est constatée dans les ontologies obtenues par la transformation d'un thésaurus. Ceci pose problème lorsqu'un utilisateur ou une application choisit d'explorer l'ontologie par une navigation de haut en bas. Le grand nombre de concepts du premier niveau rend le départ de sa navigation délicate. Par exemple, le niveau hiérarchique le plus générique de l'ontologie extraite du thésaurus IAU à cette étape de la transformation contient 1132 concepts.

Nous proposons donc l'ajout de niveaux hiérarchiques plus génériques qui facilitent la navigation dans l'ontologie. D'autre part, nous proposons la définition de concepts génériques (ou types abstraits) permettant de caractériser les concepts. Un concept générique ou abstrait fait référence à une notion abstraite et n'admet pas d'instance. Il est soit un véritable concept du domaine, soit un concept ajouté pour structurer la représentation [Furst 2004]. Dans [Soergel 2004], les concepts génériques sont définis à partir d'un schéma de catégorisation de haut niveau existant dans le domaine. Les concepts du plus haut niveau de l'ontologie sont liés manuellement aux concepts de ce schéma. Ce procédé ne peut pas être appliqué à tous les domaines car de tels schémas n'existent pas toujours. De plus, il demande un travail manuel à l'expert qui doit affecter les milliers de classes de l'ontologie à l'une des centaines de classes du schéma. Nous proposons donc une autre approche plus automatisée.

4.1.3.1 Premier niveau de généralisation : tête et expansion des syntagmes

Pour créer un premier niveau d'abstraction, les concepts sont regroupés à partir de la tête des termes de leur label. Cette approche est suivie dans OntoLearn [Velardi 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 5.8). 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 5.9). Ce mécanisme permet de créer un nouveau premier niveau de la hiérarchie contenant un nombre plus réduit de concepts.

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

$\Rightarrow c_1$ « est une sous-classe de » $F(tete(F^{-1}(c_1)))$
et c_2 « est une sous-classe de » $F(tete(F^{-1}(c_1)))$

(R6)

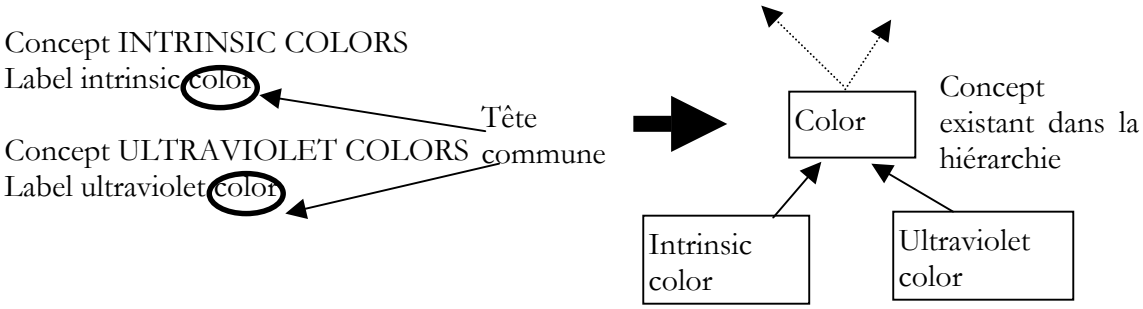


Figure 5.8 Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie

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

=> 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)

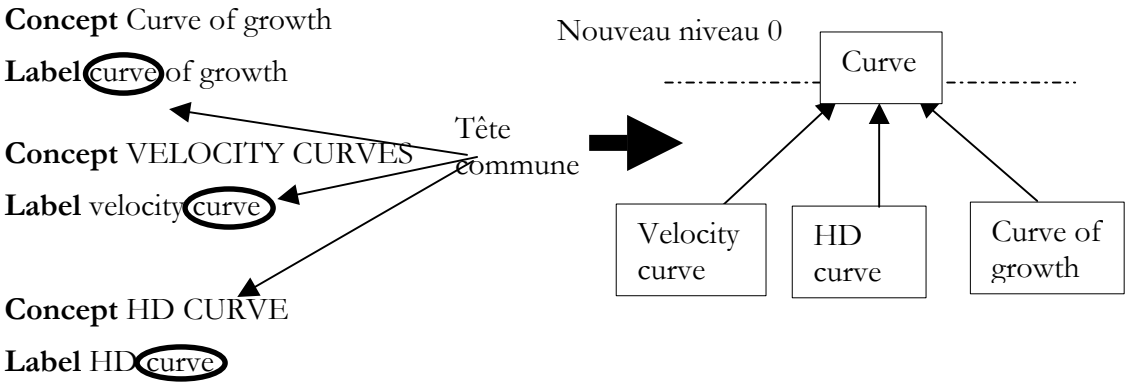


Figure 5.9 Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

4.1.3.2 Deuxième niveau de généralisation : types abstraits

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

Si c $\Leftrightarrow$ sw avec sw$\in${synsets_WordNet}

=> c « est sous-classe de » ta

Avec ta (type abstrait) le plus spécifique hyperonyme de sw

(R8)

- *Définition des types abstraits*

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

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

concepts de l'ontologie, un mécanisme de désambiguïsation est mis en place. Il prend en compte quatre éléments :

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

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

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

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

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

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

- *Association des concepts aux types abstraits*

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

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

Dans le chapitre 7, la liste des types abstraits ainsi extraits pour notre cas d'application est présentée. Un exemple d'un tel type est :

"instrumentation": an artifact (or system of artifacts) that is instrumental in accomplishing some end.

4.2 Détection des relations associatives

La deuxième étape dans la formalisation de la structure de l'ontologie vise à définir des relations associatives entre concepts de l'ontologie. Ces relations sont tout d'abord extraites des relations du thésaurus ; cette étape est présentée dans la section 5.2.1. De nouvelles relations entre les concepts sont ensuite extraites à partir de l'analyse du corpus de référence ; cette étape est présentée dans la section 5.2.2.

4.2.1 Spécification de relations entre types abstraits

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

Soient ta_1 et ta_2 deux types abstraits avec $ta_1 \in C_{Onto}$ et $ta_2 \in C_{Onto}$

Soient $r,r' \in R_{Onto}$ avec $\sigma_{ROnto}: R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine

Si $r'(c_1,c_2)$ avec $c_1 \in C_{Onto}$ et $c_2 \in C_{Onto}$ et c_1 « est sous-classe de » ta_1 et c_2 « est sous-classe de » ta_2 et $G^{-1}(r')$=« est lié à »

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

(R9)

4.2.1.1 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.

4.2.1.2 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 5.1. 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 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. La matrice plus détaillée du cas d'étude est présentée dans le chapitre 7.

Type abstrait	Instrumentation	Property	Natural object
Instrumentation	Part of/ has Part excludes	Observes	Observes
Property	Observed/measured by	Part of/ has Part excludes	Is a property of
Natural object	Observes/measures by	Has a	Part of/ has Part

Tableau 5.1 Extrait de la matrice des relations entre types abstraits

4.2.1.3 Association des relations vagues du thésaurus et des relations entre types

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 [Soergel 2004]. Les relations entre concepts sont en effet établies à partir de l'analyse des relations du thésaurus et de la définition de patrons permettant de retrouver les relations sémantiques spécifiées dans l'ensemble du corpus. Plutôt que d'avoir à spécifier individuellement les relations vagues (dans le thésaurus) entre termes, l'expert doit seulement valider ou invalider les propositions qui lui sont faites sur la base de l'analyse du corpus et des relations entre les types abstraits. Ainsi, l'analyse que nous mettons en place facilite le travail de l'expert.

4.2.2 Détection de nouvelles relations associatives

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

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

La première prend en compte les termes co-occurrant fréquemment autour des labels de concepts de l'ontologie.

La seconde repose sur l'analyse distributionnelle réalisée par le module UPERY de Syntex décrit dans le chapitre 2 section 2.6.2. Ce type d'analyse consiste à rapprocher des syntagmes en fonction de la ressemblance de leur contexte. Les syntagmes déduits de l'analyse syntaxique sont rapprochés s'ils sont formés autour de la même relation et des mêmes têtes et queues. Par exemple, en considérant les syntagmes « star », « galaxy », « star mass » et « galaxy mass », les syntagmes « star » et « galaxy » sont rapprochées par le contexte « mass ». UPERY permet de

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

Lorsqu'un label apparaît dans le contexte d'un concept ou 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 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 5.1). 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 $\sigma_{ROnto}: R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine

Si $r'(c_1,c_2)$ décelée par l'analyse du corpus avec $c_1 \in C_{Onto}$ et $c_2 \in C_{Onto}$

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

(R10)

5 Mise à jour de l'ontologie

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.

5.1 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 (cf section 3.2 de ce chapitre). 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 5.10 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 chapitre 7 pour les évaluations).

```
column density
high resolution
globular cluster
white dwarf
binary system
soft X ray
time scale
orbital period
stellar population
power law
absorptance line
line emission
active region
```

Figure 5.10 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.

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

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

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 5.11 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 (cf chapitre 7 pour les évaluations).

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

$\Rightarrow t \in L_{COnto}$

(R12)

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

Figure 5.11 Termes spécifiques de l'astronomie non présents dans le thésaurus

5.2 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 $\sigma_{ROnto}: R_{Onto} \rightarrow C \times C$ et $r(ta_1,ta_2)$ avec $G^{-1}(r)$ spécifiés dans le domaine

$$=> r' \in R_{Onto} \text{ avec } G^{-1}(r') \in G^{-1}(r)$$

(R13)

Soient $t \in L_{corpus}$ à ajouter dans L_{COnto} ,

Si $tete(t) \in L_{COnto}$ et $queue(t) \notin L_{Conto}$

$$=> t \in L_{COnto} \text{ avec } F(t)=c \text{ et } c \text{ « sous-classe de » } F(tete(t))$$

(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.

6 Conclusion

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) et la mise à jour de la connaissance à partir d'un corpus de référence. Les procédés sont simples à mettre en œuvre et permettent d'extraire une ontologie légère. Ils nécessitent une validation par un expert du domaine, mais le travail qui lui est demandé est allégé par la proposition d'éléments à chacune des étapes. L'expert est moins sollicité que dans les approches proposées dans [Soergel 2004] [Wielinga 2001] 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.

Une contribution importante de ce travail est la proposition permettant de déceler puis de labelliser les relations associatives entre concepts. Elle repose sur la notion de type abstrait qui sont des concepts de haut niveau d'abstraction. La définition de relations sémantiques, validée par des experts, est rapide compte 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 ».

Nos contributions décrites dans ce chapitre ont été diffusées et reconnues par la communauté scientifique nationale et internationale. Les résultats sur l'extraction de concepts et de relations entre concepts, non présents dans le thésaurus IAU et détectés par nos méthodes, ont été présentés lors de la réunion de l'International Virtual Observatory Alliance (septembre 2005, Espagne), ainsi qu'à la journée Ontologie dans le cadre du projet MDA (octobre 2005, Strasbourg). Une présentation générale de la méthode de transformation du thésaurus IAU en une ontologie légère et les utilisations possibles d'une telle ressource sera effectuée à la XXVIième assemblée générale de « International Astronomical Union » (IAU) en Août 2006 (Prague).

Chapitre 6
Adéquation d'une ontologie à un corpus
Méthodologie et mesures de comparaison

1 Introduction

Le choix des ontologies sur lesquelles un SRI peut s'appuyer est une question non triviale. Dans le contexte de la RI, l'utilisation d'ontologies a pour but d'aider le système dans son rôle d'intermédiaire entre le corpus de documents et l'utilisateur. La connaissance représentée doit permettre au système d'interpréter d'une part le contenu des documents et d'autre part, le besoin de l'utilisateur. Pour cela, une des utilités des ontologies est de spécifier le contexte dans lequel s'effectue la recherche à travers la modélisation de la connaissance abordée dans le corpus. Si certaines approches considèrent que seule une connaissance sur le domaine est nécessaire [Mihalcea 2000] [Vallet 2005], nous considérons au contraire que pour atteindre ces objectifs, l'ontologie doit non seulement couvrir le domaine de connaissance lié au corpus, mais également apporter des éléments de connaissance utiles pour la compréhension du corpus considéré (couverture du lexique de l'ontologie, organisation des concepts dans la structure de l'ontologie). Comme nous l'avons présenté dans le chapitre 3 section 3, le choix d'une telle ontologie peut être réalisé à partir d'analyses quantitatives. Ce type d'analyse consiste à évaluer la pertinence du choix d'une ontologie pour la réalisation d'une tâche. Ainsi, dans le contexte général de la RI, ce type d'analyse est adapte et peut être réalisé suivant différentes perspectives.

La première perspective consiste à considérer un procédé de RI à réaliser (indexation sémantique, désambiguïsation de requête, exploration de corpus, …) et à comparer les résultats obtenus pour cette tâche grâce à l'utilisation de différentes ontologies. Cette approche est longue à mettre en place et présente de nombreuses limites quant aux conclusions qui peuvent être tirées de l'analyse. Cette analyse est dépendante du corpus sur lequel s'effectue la tâche ainsi que des performances du système choisi pour réaliser la tâche sur ce corpus. Il est en effet difficile de prédire les performances d'un système sur un corpus car elles dépendent des spécificités de la collection et des requêtes. En prenant par exemple le système de RI SMART [Salton 1971] dont les performances des différentes versions ont été comparées sur les 8 premières éditions de la tâche ad-hoc de la campagne TREC[1], on s'aperçoit que pour chaque version du système les performances varient en fonction du corpus et des requêtes données à valider pour l'année de la campagne [Buckley 2000]. Buckley montre de plus que la meilleure version du système sur la collection TREC-7 n'est pas forcément la meilleure version sur la collection TREC-8. Si l'on ajoutait une ontologie à ce système, les seules conclusions auxquelles on pourrait aboutir seraient l'intérêt pour ce système sur la collection considérée. Il serait donc difficile de généraliser son intérêt pour une tâche de RI sans biaiser les résultats par la performance du système sur le corpus considéré.

Ainsi, il est nécessaire de se placer dans un contexte plus général. Nous souhaitons évaluer l'ontologie non pas par rapport aux résultats qu'elle permet d'obtenir sur une tâche donnée de RI mais par rapport à la connaissance qu'elle offre pour préciser le contexte d'interprétation du corpus pour n'importe quelle tâche. L'analyse de l'ontologie porte alors sur l'adéquation entre la connaissance de l'ontologie et les données du corpus. Dans l'ensemble des tâches de RI, l'ontologie doit, en effet, permettre d'expliciter le sens des informations du corpus. Dans le cadre de l'indexation sémantique, l'ontologie est utilisée pour extraire des descripteurs faisant référence aux entités conceptuelles traitées dans les documents. Dans le cadre de la re-formulation de requête ad-hoc et de l'appariement document/requête, les ontologies servent à identifier les concepts des documents désignés par les termes de la requête. Dans le cadre de l'exploration, les ontologies permettent de regrouper les documents abordant les mêmes concepts. Préciser le contexte d'interprétation du corpus est donc le premier procédé auquel une

[1] Text REtrieval Conference, trec.nist.gov

ontologie doit répondre en RI. Afin de mettre en place ce type d'analyse, nous proposons une méthodologie et des mesures visant à évaluer l'adéquation entre un corpus et une ontologie.

Dans la littérature, les approches visant à évaluer l'adéquation entre un corpus et une ontologie limitent leur analyse aux hiérarchies de concepts (voir chapitre 3 section 3.2.2). La méthodologie que nous proposons vise à prendre en compte non seulement les relations hiérarchiques mais aussi les relations non-taxonomiques ou associatives entre les concepts. Ce type de relations permet de préciser le sens des concepts dans l'ontologie et donc d'analyser l'adéquation à partir d'un niveau sémantique plus élevé. Contrairement aux approches existantes, la priorité de la méthodologie est d'évaluer l'adéquation entre l'ontologie et le corpus sans biaiser l'analyse par l'utilisation de ressources extérieures telles que WordNet. L'ontologie doit, à elle seule, représenter le contexte du corpus. La méthodologie que nous proposons a pour but de fournir un cadre général pour l'évaluation de l'adéquation entre une ontologie et un corpus en intégrant les aspects développés dans les travaux existants et en ajoutant les analyses nécessaires pour l'évaluation d'une ontologie légère.

Ce chapitre est organisé de la façon suivante. Tout d'abord, la section 2 présente la méthodologie ainsi que les différents critères à prendre en compte pour évaluer l'adéquation entre un corpus et une ontologie. La section 3 décrit les méthodes permettant de mettre en œuvre la méthodologie.

2 Méthodologie

Notre méthodologie d'évaluation d'adéquation s'appuie sur un corpus et une ontologie.

Les **corpus** considérés contiennent des granules documentaires. Un corpus peut être formé de documents non structurés, semi-structurés ou structurés. Il peut contenir des documents ne comportant que du texte comme des documents intégrant des images légendées textuellement. Comme nous l'avons vu au chapitre 4, le corpus est analysé à partir de son lexique L_{Corpus} contenant l'ensemble des syntagmes extraits par un analyseur syntaxique.

Les **ontologies** considérées sont des ontologies dites légères. En effet, les ontologies lourdes définies à base d'axiomes nécessitent de lourds efforts de conception ; elles sont donc peu nombreuses. De plus, par leur degré de formalisation élevé, elles sont plus difficilement réutilisables. Nous nous focalisons donc sur les ontologies légères. Comme nous l'avons vu dans le chapitre 4, une ontologie est formalisée à partir d'une structure S $\{I, C, R , \leq^C, \sigma_R, \sigma_A, \}$ et d'un lexique L $\{L^C,L^R,F,G\}$. Dans le cas de notre étude, nous nous limitons à des ontologies ne contenant pas d'instances ni de relations d'attributs, mais seulement des concepts, comme cela est fait dans la majorité des travaux du domaine [Maedche 2002a] [Desmontils 2002]. Ainsi l'ontologie est formalisée par :

la structure S $\{C, R , \leq^C, \sigma_R, \}$ et le lexique L $\{L^C,L^R, F,G\}$

Le niveau lexical couvre tous les termes ou labels définis pour désigner les concepts et les relations. Le niveau conceptuel représente les concepts et la sémantique qui leur est associée à partir des relations conceptuelles qui les lient. Notons que le lexique et la structure d'une ontologie lourde peuvent donc être également évalués par cette méthodologie.

La section 2.1 présente les critères qui doivent être pris en compte pour l'évaluation de l'adéquation. La section 2.2 présente les différentes étapes qui permettent d'évaluer ces critères. Ces différentes étapes seront reprises dans la section 3 pour expliciter les fonctions qui permettent de les mettre en œuvre.

2.1 Critères de l'adéquation

L'objectif de la méthodologie est d'évaluer dans quelle mesure l'ontologie permet de préciser et d'expliciter le contexte associé au corpus. Dans le cadre de l'évaluation de son adéquation à un corpus, nous distinguons différents critères en fonction du fait qu'ils permettent de décider directement de la non adéquation de l'ontologie, ou qu'ils permettent de mesurer cette adéquation.

2.1.1 Critères de filtrage d'ontologies

Un premier ensemble de critères porte sur une appréciation préalable des caractéristiques de l'ontologie et du corpus. Ces critères ont pour objectif d'éliminer une ontologie ne répondant pas à des caractéristiques fondamentales d'adéquation.

La définition de ces critères est motivée par la conclusion présentée dans [Desmontils 2002] selon laquelle l'adéquation entre le corpus et l'ontologie n'est pas suffisante si les domaines traités par l'ontologie ne couvrent pas ceux du corpus.

Les critères que nous retenons portent sur les caractéristiques suivantes :

- les domaines abordés dans les ressources,
- l'époque à laquelle les ressources ont été élaborées,
- la validité de l'ontologie.

Les **domaines de connaissances** abordés dans l'ontologie et dans le corpus doivent tout d'abord être identifiés. Les ressources (ontologie et corpus) doivent être considérées à partir du ou des domaines principalement traités ainsi que les domaines connexes qui y sont associés. Les domaines doivent ensuite être comparés de façon à établir si l'ontologie couvre tous les domaines de connaissances abordés dans le corpus. Une ontologie ne couvrant qu'une partie des thématiques du corpus ne représentera que partiellement son contexte. Par exemple, une ontologie de l'astronomie n'intégrant pas le domaine de la physique ne pourra pas être utilisée pour préciser le contexte d'un corpus constitué d'articles publiés dans une revue généraliste du domaine, car l'astronomie intègre les éléments de la physique et les informations du corpus couvrant ces éléments ne pourront pas être spécifiés dans le contexte de l'ontologie. Une ontologie intégrant des domaines non-traités dans le corpus reste exploitable dans la mesure ou elle représente la connaissance précisant le contexte du corpus, bien qu'elle soit constituée de connaissances supplémentaires. Ce critère reste un peu superficiel et subjectif.

La **date de création des ressources** (ontologie et corpus) est également un facteur à prendre en compte. Dans le cas d'ontologies de domaines scientifiques par exemple, la connaissance évolue très rapidement. Ceci est notamment le cas dans le domaine de l'astronomie où de nombreuses découvertes enrichissent annuellement le domaine (exemple des planètes telluriques hors de notre système solaire découvertes en 2004). Il est donc nécessaire de choisir une ontologie qui intégrera la connaissance du domaine au moment de la création des documents du corpus.

La **validité de l'ontologie** à travers l'exactitude des connaissances qu'elle représente et la formalisation de sa représentation sont aussi des facteurs à prendre en compte. L'ontologie doit en effet représenter la connaissance vérifiée dans le domaine. Elle doit aussi être formalisée correctement. Dans ce cas-là, les critères tels que la consistance, la complétude, la concision, l'expansibilité et la sensibilité de l'ontologie aux changements décrits dans [Gomez-Perez 1999] sont pris en compte.

2.1.2 Critères d'analyse de l'adéquation

Un deuxième ensemble de critères porte sur l'adéquation des ressources par rapport à leur contenu. L'analyse porte alors sur la comparaison de la connaissance représentée dans l'ontologie avec les données contenues dans le corpus. Ces critères nécessitent une analyse approfondie des deux ressources à partir des deux niveaux sémiotiques de l'ontologie : son lexique et sa structure. Le lexique de l'ontologie regroupe les termes référençant les concepts, permettant ainsi de les identifier dans le corpus. La structure de l'ontologie permet de spécifier le sens des concepts et donc d'interpréter la sémantique qui leur est associée. Cette sémantique doit être adaptée au contexte du corpus.

L'analyse de l'adéquation du contenu des ressources porte alors sur

- l'adéquation lexicale qui permet d'identifier les concepts du corpus,
- l'adéquation conceptuelle qui permet de capturer la sémantique associée aux concepts.

2.1.2.1 Adéquation lexicale

L'adéquation lexicale consiste à comparer le lexique des deux ressources. Le lexique défini dans l'ontologie doit permettre d'identifier l'ensemble des concepts référencés dans le corpus. Le corpus étant constitué de termes, et afin d'accéder aux concepts qu'ils désignent, il est indispensable que le lexique de l'ontologie permette d'identifier les concepts mentionnés dans le corpus.

Les critères d'adéquation sont alors :

- la présence des termes du corpus comme labels de concepts dans l'ontologie,
- la détection des concepts désignés par les labels dans le cas où les labels retrouvés dans le corpus se rapportent à différents concepts.

2.1.2.2 Adéquation conceptuelle

L'organisation des concepts retrouvés dans le corpus doit également refléter le sens des concepts dans le corpus. Ce critère se rapporte à l'adéquation conceptuelle. L'ontologie doit présenter de la connaissance utile sur les concepts du corpus et prendre le point de vue du corpus, c'est-à-dire décrire la sémantique associée aux concepts dans le sens où ils sont référencés dans les documents. Afin d'illustrer ce critère, nous faisons référence aux extraits d'ontologies schématisés dans la figure 6.1. Les concepts sont représentés par des rectangles. Les labels des concepts se trouvent dans les rectangles, le label principal étant en gras. Les relations sont représentées à l'aide de flèches, la légende associée à la figure fournit le label des relations et leurs caractéristiques.

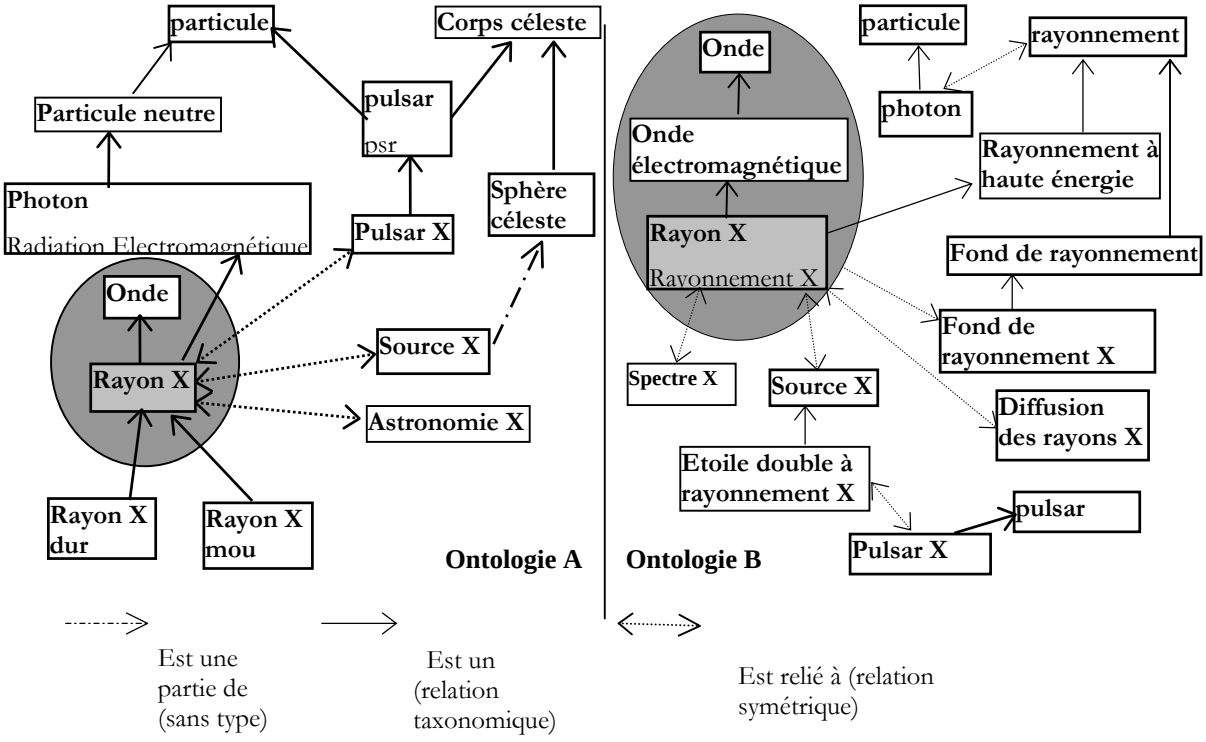


Figure 6.1 Extraits de deux ontologies de l'astronomie

L'ontologie B de la figure 6.1 est extraite du site[2] et l'ontologie A a été créée à partir du thésaurus IAU[3]. Les deux ontologies représentent une conceptualisation possible du domaine des Rayons X. Elles ont toutes deux été validées par des experts du domaine. Cependant, des différences notables les distinguent. Par exemple, un *Rayon X* est défini dans l'ontologie B comme étant un sous-concept d'une *Onde électromagnétique* lui-même sous-concept d'une *Onde*. Dans l'ontologie A, un *Rayon X* est défini directement comme étant un sous-concept de *Onde*. De plus, dans l'ontologie A, un *Rayon X* est un sous-concept de *Particule* alors que dans l'ontologie B, un *Rayon X* est lié indirectement à *Particule* à partir des concepts *Radiation* et *Photon*. En partant de ces considérations, il est intéressant de noter que les représentations ne sont pas issues des mêmes choix de conceptualisation.

Un critère d'adéquation corpus-ontologie est donc le choix d'une conceptualisation qui précise le contexte du corpus. Une conceptualisation représentative du corpus doit prendre en compte le degré de spécialisation-généralisation des concepts utiles pour la représentation du contexte associé au corpus. En reprenant l'exemple précédent, le concept *Onde électromagnétique* aura sa place dans l'explicitation du contexte du corpus si ce concept est référencé dans les documents. Il permettra en effet de préciser le contexte d'un document traitant des *Rayons X* par rapport à un document traitant des *Ondes*.

La conceptualisation doit également permettre d'interpréter le sens des concepts retrouvés dans les documents à partir du contexte général décrit dans l'ontologie. La représentation d'un concept dans l'ontologie doit en effet permettre de préciser le rôle du

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

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

concept dans l'explication du contexte du corpus. Un concept lié sémantiquement à plusieurs autres concepts retrouvés dans les documents explicitera son contexte, alors qu'un concept isolé, ou lié à peu de concepts du corpus, ne le permettra pas. A partir de l'exemple précédent, le choix de conceptualisation du concept *Particule* doit prendre en compte les concepts du corpus. Dans le cas de l'ontologie B, si aucun des concepts liés directement ou indirectement à *Particule* n'est présent dans le corpus, sa représentation dans l'ontologie ne sera pas adaptée à la spécification du contexte du corpus et le choix de conceptualisation de l'ontologie A est préférable.

Dans la perspective d'évaluer l'adéquation conceptuelle, les relations sémantiques entre les concepts de l'ontologie retrouvées dans le corpus doivent pouvoir être évaluées. Ces liens doivent être établis à travers les relations et leurs caractéristiques. Les caractéristiques permettent d'évaluer dans quelle mesure les concepts sont liés sémantiquement. Considérons par exemple les concepts *Rayon X* et *Pulsar X* dans l'ontologie A et B retrouvés dans un corpus. Par les caractéristiques de la relation *est un*, l'ontologie A permet de préciser que *Rayon X* et *Pulsar X* sont tous deux subsumés par *Particule*. Ce sont donc deux concepts spécifiant le concept *Particule*. Dans l'ontologie B, *Rayon X* est tout d'abord lié à *Source X* par une relation non transitive qui a un sous-concept *Etoile double à rayonnement X* lui-même lié par une relation non-transitive à *Pulsar X*. Le lien sémantique entre les concepts *Rayon X* et *Pulsar X* est donc moins fort dans l'ontologie B que dans l'ontologie A. De plus, pour une même relation sémantique, le degré sémantique de la relation peut être différent en fonction des concepts qu'elle relie. Par exemple, le niveau de spécification de la relation *est un* doit être considéré. Dans le cas de l'ontologie A, la relation *est un* liant les concepts *Rayon X* et *Onde* implique un niveau de spécification moindre que la relation est un liant les concepts *Rayon X* et *Onde électromagnétique*.

L'adéquation conceptuelle doit donc être analysée à partir de :

- la représentativité des relations taxonomiques,
- la représentativité des relations non-taxonomiques.

2.2 Etapes de la méthodologie

Afin de mettre en œuvre l'analyse de l'adéquation entre un corpus et une ontologie à partir des différents critères présentés dans la section précédente, nous proposons une méthodologie comportant quatre étapes principales. L'ensemble de ces étapes est présenté dans la figure 6.2. Les trois premières étapes permettent d'analyser l'adéquation à différents niveaux. La première étape vise à analyser les caractéristiques préalables du corpus et de l'ontologie. Les deux étapes suivantes portent sur l'analyse du contenu des ressources [Hernandez 2004a]. La deuxième étape correspond ainsi à l'analyse lexicale et la troisième étape à l'analyse conceptuelle. Finalement, la dernière étape repose sur l'interprétation des résultats et le choix de l'ontologie.

La spécification d'une méthodologie permet d'analyser chacun des critères indépendamment et d'identifier leur impact sur l'adéquation. Contrairement aux approches présentées dans la littérature [Brewster 2004], [Rottenburger 2002], [Desmontils 2002], le cadre d'évaluation de l'adéquation se veut généraliste et vise à isoler chacune des étapes afin d'analyser la qualité de chacune d'entre elles.

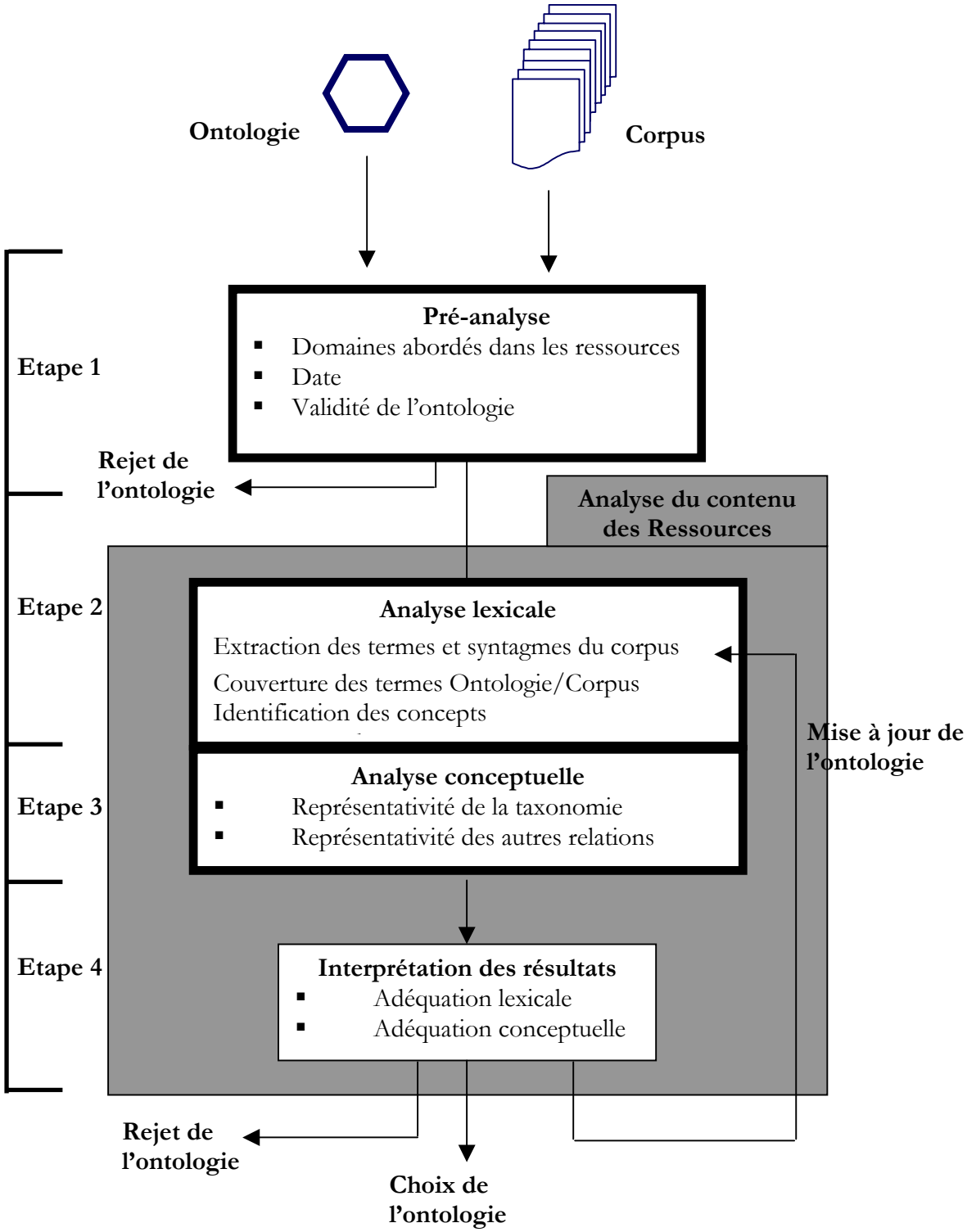


Figure 6.2 Etapes de la méthodologie

2.2.1 Pré-analyse

La pré-analyse évalue l'adéquation entre le corpus et l'ontologie à partir des caractéristiques des ressources. Cette étape vise à éliminer une ontologie dont les caractéristiques ne seraient pas adaptées au corpus. Elle n'est intégrée dans aucune des méthodes trouvées dans la

littérature. Les caractéristiques de cette analyse peuvent être extraites de la documentation disponible sur l'ontologie et des informations générales sur le corpus.

2.2.1.1 Domaines abordés dans les ressources

La documentation de l'ontologie peut spécifier les domaines représentés ainsi que ses limites comme, par exemple, l'intégration de domaines connexes. Le domaine couvert par l'ontologie peut également être précisé à travers les sources qui ont permis son élaboration. Une ontologie élaborée à partir d'un thésaurus peut indiquer qu'un lourd effort a été mené pour intégrer les termes consensuels du domaine. Une ontologie réalisée à partir de comptes rendus d'experts indique que la connaissance tacite du domaine a été capturée. Le domaine est également précisé par la tâche à laquelle doit répondre l'ontologie lors de sa création. Une ontologie élaborée pour le traitement de dossiers médicaux pourra intégrer en plus des maladies, diagnostics et traitements donnés par les médecins et des informations sur les malades.

Les domaines traités dans le corpus peuvent être indiqués par les raisons qui ont motivé sa constitution. Par exemple, les thématiques d'un corpus constitué d'articles d'une revue scientifique seront celles de la revue. Les domaines peuvent également être indiqués par la nature des documents du corpus. Un corpus contenant des relevés d'observations télescopiques couvrira le domaine des éléments observés ainsi que les paramètres observables à partir du télescope.

Lorsque les domaines des ressources peuvent être clairement identifiés, une ontologie est rejetée si elle ne couvre pas l'ensemble des domaines du corpus.

L'adéquation pour ce critère est formalisé par la règle R1 :

si couverture insuffisante => rejet de l'ontologie **(R1)**

2.2.1.2 Date des ressources

Dans le cas où le domaine de connaissance est stable, ce critère n'est pas à considérer. Cependant, il est important dans le cadre des domaines scientifiques pour lesquels la connaissance évolue constamment.

La période pour laquelle une ontologie capture la connaissance d'un domaine peut être extraite des documentations associées à la ressource. La date de création des documents utilisés pour construire une ontologie à partir de textes indique par exemple cette période. Dans le cas du corpus, cette période correspond, par exemple, à l'année de publication des articles composant le corpus.

Si les ressources ne datent pas d'une même période, l'ontologie est rejetée.

L'adéquation pour ce critère est formalisée par la règle R2 :

si période(corpus) ≠ période(ontologie)=> rejet de l'ontologie **(R2)**

2.2.1.3 Validité de l'ontologie

La validité de l'ontologie peut tout d'abord être analysée à partir de la fiabilité de ses auteurs. Les auteurs doivent être considérés à la fois par rapport à leur niveau d'expertise sur le domaine et à leur expérience dans la conception d'ontologies. Une ontologie créée par un

amateur dans le cadre de ses loisirs est moins fiable qu'une ontologie créée par des experts du domaine et des concepteurs d'ontologies. Ce sous-critère reste subjectif.

L'ontologie est ensuite analysée par rapport à la validité de sa représentation. Des outils tel que ODEval peuvent être utilisés pour valider la consistance de la taxonomie de l'ontologie [Corcho 2004]. Cet outil s'applique aux ontologies formalisées en RDF(S), DAML+OIL et OWL et identifie l'inconsistance et la redondance dans la représentation. Une ontologie non fiable est rejetée.

L'adéquation pour ce critère est formalisée par la règle R3 :

si validité(ontologie) incorrecte => rejet de l'ontologie **(R3)**

Une ontologie rejetée doit être modifiée avant de mesurer son adéquation à un corpus par les étapes suivantes.

2.2.2 Analyse lexicale

L'analyse lexicale consiste à comparer le lexique (ou ensemble de termes) de l'ontologie et le lexique de la collection. Cette étape est cruciale car elle permet d'évaluer la présence des termes permettant de désigner les concepts et d'identifier les concepts référencés dans le corpus. La comparaison implique la projection des termes de l'ontologie sur le corpus et la projection des termes du corpus sur l'ontologie.

Afin que cette étape soit mise en place, les termes et expressions du corpus (L_{corpus}) doivent être extraits. L'extraction peut reposer soit sur une extraction statistique soit sur une extraction syntaxique. Toutes les techniques d'extraction décrites dans la section 2.4 du chapitre 2 peuvent être utilisées. Les labels des concepts de l'ontologie sont ensuite recherchés dans l'ensemble des termes extraits. Le nombre de termes retrouvés reflète l'adéquation lexicale du corpus (R4). Les termes représentatifs du corpus sont également calculés à partir des termes extraits du corpus. Cette étape vise à extraire les termes importants du corpus. Le nombre de termes représentatifs du corpus retrouvés dans l'ontologie indique l'adéquation lexicale de l'ontologie (R5). La phase finale de l'analyse lexicale consiste à déterminer les concepts représentés par les labels de l'ontologie retrouvés dans le corpus. Dans le cas où un label désigne plusieurs concepts, un mécanisme de désambiguïsation du sens du label doit être mis en place afin d'identifier le concept effectivement référencé. Le mécanisme choisi doit permettre d'identifier correctement les concepts afin de permettre l'analyse conceptuelle. Le nombre de labels pour lesquels l'ambiguïté est levée indique l'ambiguïté lexicale de l'ontologie (R6).

Cette analyse lexicale aboutit aux règles suivantes :

$card(L^C \cap L_{corpus}) \Rightarrow$ adéquation lexicale(corpus) (R4)

$card(termes\_représentatifs(L_{corpus}) \cap L^C)$ => adéquation lexicale(ontologie) (R5)

card { l/ l∈ $L^C \cap L_{corpus}$, ∃ Amb / Amb(l)=c avec c∈ F(l)) => ambiguïté lexicale(ontologie) (R6)

Les méthodes permettant d'évaluer ces règles sont décrites dans la section 3.1.

2.2.3 Analyse conceptuelle

L'analyse conceptuelle vise à déterminer si l'organisation des concepts au sein de la structure de l'ontologie apporte de la connaissance utile pour l'interprétation des données présentes dans une collection documentaire. Elle se focalise sur la comparaison des liens sémantiques entre les concepts dans l'ontologie avec les liens sémantiques entre les labels des concepts dans le corpus. Cette analyse permet d'établir l'adéquation entre un corpus et deux ontologies représentant une conceptualisation différente ou bien d'évaluer si l'ontologie permet de spécifier le contexte d'un concept. Cette analyse doit prendre en compte à la fois les relations taxonomiques de l'ontologie et les relations non-taxonomiques.

Elle repose sur deux étapes. Le calcul de l'adéquation conceptuelle de chacun des concepts de l'ontologie établit si le concept est correctement situé dans la structure de l'ontologie (R7). Ce calcul réalisé globalement au niveau de l'ontologie permet de quantifier l'adéquation de l'ontologie (R8).

∀c ∈ C_{onto} / adéquation_conceptuelle(c)<seuil => mauvais positionnement du concept dans l'ontologie (R7)

{adéquation_conceptuelle(c)} pour c ∈ C_{onto} => adéquation_conceptuelle (Ontologie) (R8)

Les mesures permettant d'évaluer cette adéquation sont présentées dans la section 3.2

2.2.4 Interprétation des résultats

La dernière phase est celle de l'interprétation des résultats. Cette phase doit permettre de conclure sur l'adéquation de l'ontologie au corpus à partir de l'impact des différents critères. Elle doit permettre d'évaluer la pertinence des méthodes choisies pour effectuer les différentes analyses en permettant d'accéder aux éléments traités (concepts, labels des concepts, relations conceptuelles, documents du corpus).

Les trois premières règles permettent directement de décider du rejet éventuel de l'ontologie.

Pour chacune des autres règles, des seuils d'adéquation sont fixés empiriquement pour conclure sur le rejet de l'ontologie. Ces règles peuvent également être utilisées pour déclencher des mises à jour qui permettraient d'améliorer l'adéquation de l'ontologie au corpus. Par exemple, une faible adéquation lexicale du corpus pourra être améliorée par l'ajout de nouveaux labels de concepts ou de nouveaux concepts. Ces aspects de mises à jour sont étudiés dans le chapitre 5.

Dans la section suivante, nous présentons les méthodes permettant de mettre en œuvre l'ensemble des étapes de la méthodologie en les positionnant par rapport à l'état de l'art.

3 Evaluer l'adéquation du contenu des ressources

Des méthodes et des mesures sont nécessaires pour effectuer l'analyse lexicale et l'analyse conceptuelle. Contrairement aux approches présentées dans [Brewter 2004] et [Rottenburger 2002], l'objectif des méthodes proposées est de permettre d'évaluer l'adéquation en minimisant le travail nécessitant une intervention humaine.

3.1 Analyse lexicale

Dans la littérature, les termes sont extraits du corpus manuellement [Rottenburger 2002], syntaxiquement [Desmontils & Jacquin, 2002] ou statistiquement [Brewster 2004]. Les labels de l'ontologie sont ensuite recherchés dans l'ensemble des termes extraits. Ces approches se distinguent par les méthodes utilisées pour sélectionner les termes. Cette étape d'extraction de termes représentatifs est ignorée dans [Rottenburger 2002] où les expressions sont extraites du corpus manuellement par l'intuition selon laquelle elles représenteraient un concept. Dans [Desmontils 2002], [Brewster 2004], le lexique extrait est uniquement celui jugé représentatif du corpus, à partir des balises HTML encadrant les termes et de leur fréquence dans le premier cas et à partir de l'analyse de la sémantique latente dans le deuxième. Aucune méthode ne considère l'ensemble des termes extraits des corpus. Il est cependant nécessaire de considérer ces deux aspects dans le cadre de l'évaluation de l'adéquation entre un corpus et une ontologie. Tout d'abord, comme nous l'avons souligné dans la section 3 du chapitre 2, aucune méthode ne permet d'extraire exactement l'ensemble des termes représentatifs. Le fait de considérer tous les termes issus des documents permettrait alors d'identifier les termes du corpus qui auraient été ignorés par la sélection des termes représentatifs. D'autre part, les termes jugés non représentatifs du document peuvent néanmoins apporter de l'information sur le contexte du corpus et aider à préciser le sens des termes représentatifs. Il est donc important de retrouver un maximum de termes du corpus dans l'ontologie. Nous considérons ces deux aspects dans les mesures que nous proposons dans les sections 3.1.2, 3.1.3 et 3.1.4.

Bien qu'elle soit cruciale, la phase d'identification des concepts désignés par les termes du corpus n'est analysée dans aucune des méthodes de la littérature. Cette phase est simplifiée dans [Rottenburger 2002] car aucun label ne peut se rapporter à différents concepts. Dans [Brewster 2004], elle est réalisée manuellement. Une étape de désambiguïsation est présentée dans [Desmontils 2002] mais elle n'est pas évaluée indépendamment, alors qu'elle influence la qualité de l'analyse conceptuelle. Dans notre démarche et les mesures associées présentées dans la section 3.1.4, nous distinguons clairement l'étape de désambiguïsation de l'étape d'extraction lexicale, permettant ainsi une évaluation indépendante de chaque critère.

3.1.1 Extraction du lexique du corpus

Comme nous l'avons vu dans le chapitre 4, l'analyseur syntaxique de corpus Syntex [Bourigault 2000] est utilisé pour extraire les termes (mots simples ou syntagmes) du corpus. L'extraction syntaxique du lexique a été préférée à une extraction statistique. Les labels d'une ontologie étant généralement représentés à partir de syntagmes, il est primordial de faire reposer l'extraction des termes sur une technique permettant d'extraire avec précision ces éléments. L'avantage des techniques syntaxiques est de détecter les syntagmes à partir du rôle syntaxique des termes dans la phrase et non pas uniquement à partir de leur occurrence comme le permettent les techniques statistiques. Comme décrit dans la section 2.4 du chapitre 2, Syntex a, de plus, l'avantage de reposer sur un apprentissage endogène du corpus, ce qui permet de l'appliquer sur des collections variées et spécialisées. Syntex fournit également les termes sous leur forme lemmatisée. Cette forme consiste à mettre au masculin singulier les noms et adjectifs et à l'infinitif les verbes composant les syntagmes. Les variations lexicales des termes sont ainsi capturées. La forme lemmatisée correspond à la forme choisie pour la spécification des labels des ontologies. Syntex permet d'extraire des termes de plusieurs natures : noms, verbes, adjectifs,

syntagmes nominaux, syntagmes verbaux, … Le lexique du corpus est composé des syntagmes nominaux extraits du corpus par Syntex ; en effet, seul ce type de syntagmes est présent dans le lexique des concepts d'une ontologie. Syntex extrait les syntagmes maximaux ainsi que l'ensemble des syntagmes qui peuvent être simplifiés et réduits à partir de ces syntagmes. En prenant par exemple la phrase suivante « The line profiles consist of two strong velocity spikes », l'ensemble des syntagmes nominaux suivant est extrait : « line profiles » et « two strong velocity spikes », correspondant aux syntagmes maximaux, ainsi que « profile », « strong velocity spike », « velocity spike », « spike », « velocity » correspondant aux syntagmes réduits. Finalement, les syntagmes que nous considérons pour constituer le lexique du corpus (L_{corpus}) sont les syntagmes réduits et maximaux.

3.1.2 Adéquation lexicale du corpus : projection des termes de l'ontologie sur le corpus

Le pourcentage de termes ou labels de l'ontologie présents dans le corpus est calculé dans le but d'évaluer l'adéquation lexicale de l'ontologie par rapport au corpus. Les labels des concepts de l'ontologie sont recherchés dans le lexique extrait du corpus. Cette étape consiste donc à rechercher dans le corpus les termes référençant des concepts.

Ce processus est implanté, par exemple, par la recherche des chaînes de caractères communes aux deux ensembles.

3.1.3 Adéquation lexicale de l'ontologie : projection des termes du corpus sur l'ontologie

Les termes représentatifs du lexique du corpus sont extraits à partir de méthodes statistiques issues de la recherche d'information. Comme précisé dans la section 2.4 du chapitre 2, aucune méthode ne présente pour l'instant un consensus dans le domaine. L'extraction de ces termes se fait donc à partir de deux mesures ayant des propriétés différentes.

La première mesure 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.

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

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

La deuxième mesure repose sur la mesure tf.idf qui extrait les termes discriminants d'un document. Elle favorise, en effet, les termes apparaissant dans le document et n'apparaissant pas dans le reste de la collection. Cette deuxième mesure vise à extraire les termes spécifiques du corpus. 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 $tf_{terme,granule_i}$ 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

card($L_{corpus} \cap L_C$) => adéquation lexicale(ontologie) **(R5)**

card($L_{corpus} \cap L_C$) de la (R5) s'implante par

- **card({t/ globalité(t,corpus)>seuil})**
- **card({t/ spécificité(t,corpus)>seuil})**

3.1.4 Ambiguïté lexicale : identification des concepts

La dernière phase de l'analyse lexicale consiste à identifier les concepts désignés par les labels de l'ontologie retrouvés dans le corpus. Le nombre de concepts identifiés pour un terme indique l'ambiguïté lexicale. Cette phase nécessite deux étapes : le choix du label approprié par rapport aux termes du lexique du corpus et le choix du concept correspondant dans l'ontologie.

Comme pour l'extraction des concepts d'indexation présentée dans le chapitre 4, le choix du label approprié consiste à identifier le label le plus spécifique retrouvé dans une phrase. Par exemple, dans la phrase *X ray spectra are analysed*, trois syntagmes nominaux sont extraits par Syntex *X ray spectrum*, *spectrum* et *X ray*. Considérons que *X ray spectrum*, *spectrum* et *X ray* soient des labels de concepts. Chacun de ces concepts n'est pas référencé dans cette phrase. Le concept référencé est celui correspondant au label le plus long et désignant le concept le plus spécifique qui est *X Ray spectra*. La première étape de l'identification des concepts désignés par les labels consiste donc à sélectionner les labels les plus longs retrouvés dans chaque phrase.

Dans le cas où un label retrouvé dans le document correspond à plusieurs concepts, le concept effectivement référencé doit être identifié. Le mécanisme de désambiguïsation utilisé dans notre approche est relativement simple. Il prend en compte tout d'abord le contexte d'apparition du label dans la phrase à partir des autres concepts non ambigus retrouvés. Si aucun concept non ambigu n'a été retrouvé dans la phrase, le label est analysé par rapport aux concepts retrouvés dans le document. Le concept choisi est celui qui est sémantiquement le plus proche des autres concepts identifiés dans le contexte documentaire considéré. Afin d'identifier le lien sémantique entre les différents concepts candidats et les concepts du contexte, la distance entre les concepts est évaluée à partir du nombre de relations du plus court chemin séparant les concepts. Cette mesure est présentée en détail dans la section 2.1.1 du chapitre 3. Pour chacun des concepts candidats, la somme du nombre des relations est calculée. Le concept proposé est celui pour lequel cette distance est minimale. Le choix du concept est ensuite validé par un expert. Le mécanisme considéré est simple mais permet d'identifier de façon unique le concept associé à un terme. Ceci permet une analyse conceptuelle de qualité.

Ainsi l'implantation de la règle R6 consiste à évaluer la qualité de la désambiguïsation proposée. En effet, à l'issue du traitement présenté précédemment, chaque terme du lexique du corpus est associé à un concept.

card { l/ l∈ $L_C \cap L_{corpus}$, ∃ Amb / Amb(l)=c avec c∈ F(l)) => ambiguité lexicale(ontologie) (R6)

F dans R6 s'implante par :

Amb(l)=c / $\min_{c \in F(l)} imun(\sum_{c phrase \in CK} nombre\_arc(c, c_{phrase}))$

Avec CK est l'ensemble des concepts non ambigus appartenant au contexte documentaire (phrase ou granule) de l

3.2 Analyse conceptuelle

L'analyse conceptuelle vise à évaluer l'organisation des concepts au sein de la structure de l'ontologie. Cette structure doit refléter les liens entre les concepts dans le corpus.

Cette analyse est prise en compte dans les travaux de la littérature. Dans [Brewter 2004], cette analyse est réalisée en formant des regroupements de termes à partir de l'analyse de la sémantique latente. L'adéquation conceptuelle est ensuite évaluée en comparant dans quelle mesure les labels appartenant aux concepts assemblés dans un même regroupement sont plus proches dans l'ontologie que ceux appartenant à des regroupements différents. Bien que seuls les liens taxonomiques soient considérés, l'évaluation de l'adéquation conceptuelle prend en compte à la fois la distance entre les concepts dans l'ontologie et la distance entre les concepts dans le corpus. Cependant le lien sémantique entre un concept et l'ensemble des concepts retrouvés dans le corpus n'est pas pris en compte. Dans [Desmontil 2002], l'analyse conceptuelle consiste à calculer le pouvoir représentatif de chaque concept dans les pages où ils sont retrouvés et à faire la somme de ses pouvoirs pour l'ensemble des pages du site. Le pouvoir représentatif d'un concept pour une page prend en compte ses liens avec les autres concepts de l'ontologie retrouvés. Plus un concept est lié sémantiquement aux autres concepts de la page, plus il est représentatif de la page. Cette approche se rapporte au cadre général proposé pour évaluer l'adéquation conceptuelle. Cependant, la représentativité n'est calculée qu'au niveau de la page et non au niveau du corpus. Ceci présente la limite de ne pas évaluer un concept par rapport aux concepts qui peuvent être retrouvés dans les autres documents. De plus, la mesure de similarité utilisée pour calculer le pouvoir prend en compte la distance sémantique entre les concepts uniquement dans l'ontologie (les informations sur le concept dans le corpus ne sont pas considérées). La mesure considérée est celle de Wu [Wu 1994] qui, comme expliqué dans la section 2.1.1 du chapitre 3, considère uniquement les relations taxonomiques et ne permet pas d'évaluer la subjectivité dans le choix des relations. La phase d'analyse conceptuelle n'est pas détaillée dans [Rottenburger 2002]. Le poids des concepts dans les documents est pris en compte pour évaluer la couverture du corpus sur la taxonomie et la couverture de la taxonomie sur le corpus. Cependant aucun élément n'est donné pour expliciter si ce poids prend en compte l'organisation conceptuelle des concepts dans l'ontologie.

Afin de mettre en place l'analyse conceptuelle, notre approche étend celle de [Desmontil 2002] en prenant en compte l'information contenue sur les concepts dans le corpus et en évaluant la représentativité d'un concept non pas au niveau d'un granule mais au niveau du corpus. D'autre part, contrairement à l'approche précédente, nous prenons en compte tous les types de liens entre concepts. Le pouvoir représentatif d'un concept du corpus est alors évalué à partir de son lien sémantique avec les autres concepts du corpus retrouvés. La mesure s'applique

à des ontologies dans lesquelles le type des relations peut être précisé. Nous entendons par type la caractéristique qui peut être ajoutée à la relation. Des exemples de caractéristiques sont la symétrie, la transitivité, la réflexivité et la fonctionnalité. Ces caractéristiques permettent d'inférer les liens entre les concepts. Une mesure de proximité est utilisée de manière à capturer le poids de ce lien.

La proximité entre deux concepts est calculée à partir de la distance qui sépare les deux concepts par le plus court chemin jugé valide dans l'ontologie. Un chemin est jugé valide s'il ne comporte pas plus d'une relation non transitive impliquant un changement d'orientation. Chacune des relations du chemin est pondérée en fonction de son type. Cette fonction réutilise celle permettant de calculer la pondération des concepts d'indexation présentée dans le chapitre 4, section 3.1.1.3.

$$\text{Prox}_{\text{prop}}(c1,c2) = \frac{1}{1+dist_{prop}(c1,c2)} \quad \textbf{(3)}$$

Où $\text{dist}_{\text{prop}}(c1,c2)$ représente la distance pondérée en fonction des types des relations du plus court chemin jugé valide.

$\forall c \in C_{onto}$ / adéquation_conceptuelle(c)<seuil => mauvais positionnement du concept dans l'ontologie (R7)

Dans (R7) adéquation_conceptuelle(c) est implantée par :

adéquation_conceptuelle(c)= $\sum_{ci \in Ccorpus} \text{Prox}_{prop}(c,ci)$

avec C_{corpus} représentant l'ensemble des concepts de l'ontologie retrouvés dans le corpus

Au niveau de l'ontologie, l'adéquation conceptuelle est alors calculée à partir de la somme des adéquations conceptuelles de ses concepts retrouvés dans le corpus.

{adéquation_conceptuelle(c)} pour $c \in C_{onto}$ => adéquation_conceptuelle (Ontologie) (R8)

Dans (R8) {adéquation_conceptuelle(ontologie)} est implantée par

adéquation_conceptuelle(ontologie) $= \sum_{ci \in Ccorpus} adéquation\_conceptuelle(ci)$

avec C_{corpus} représentant l'ensemble des concepts de l'ontologie retrouvés dans le corpus

4 Conclusion

La méthodologie proposée vise à analyser indépendamment les différents critères évaluant l'adéquation entre un corpus et une ontologie. Elle se décompose en plusieurs étapes : une pré-analyse du corpus et de l'ontologie qui ont pour objectif d'éliminer les ontologies inadéquates, une analyse lexicale et une analyse conceptuelle des deux ressources qui ont pour but de mesurer l'adéquation d'une ontologie, permettant ainsi de sélectionner la meilleure ontologie parmi plusieurs. Contrairement aux autres méthodes de la littérature, l'approche présentée vise à proposer un cadre générique d'évaluation de l'adéquation entre un corpus et une ontologie. Différentes implantations des méthodes sont possibles.

Ce chapitre a présenté des méthodes que nous proposons pour implanter les étapes de la méthodologie. Ces méthodes permettent d'analyser avec précision les différents critères. Une des propositions originales est celle qui évalue l'adéquation conceptuelle entre les concepts. Elle repose sur la proposition d'une mesure de proximité entre concepts visant à établir la proximité sémantique entre concepts dans une ontologie légère à partir de l'information contenue dans le corpus sur les concepts et des différents types de liens entre concepts dans l'ontologie.

Cette analyse conceptuelle présente cependant des limites. Elle ne prend pas en compte les instances de concepts. De plus, elle ignore les relations d'attributs spécifiées entre instances et types de données (correspondant en OWL aux Object Properties). Une perspective d'extension des travaux serait de prendre en compte ce type d'éléments.

Nos contributions décrites dans ce chapitre ont été diffusées et reconnues par la communauté scientifique à travers différentes présentations et publications nationales et internationales [Hernandez 2004b], [Hernandez 2004a].

Partie 3

Validations

Chapitre 7
Cadre d'application : l'astronomie

1 Projet MDA

Le projet Masses de Données en Astronomie (MDA) est le cadre d'application de nos travaux. Cette section décrit les objectifs du projet, les ressources existant dans le domaine de l'astronomie utiles pour le projet, ainsi que le cadre d'évaluation que nous avons défini.

1.1 Description du projet

Nous présentons d'abord la description générale du projet, puis notre rôle dans ce projet.

1.1.1 Description générale

En astronomie, de grandes masses de données, réparties et hétérogènes, existent. Ces données sont produites par les grands observatoires au sol et spatiaux ainsi que par les grands relevés qui explorent de façon systématique le ciel ou des fractions significatives de celui-ci. Les archives en ligne des observatoires contiennent des centaines de téraoctets de données (images, spectres, séries temporelles), les catalogues répertorient des centaines de millions d'objets (plus d'un milliard pour le catalogue USNO B2), les serveurs bibliographiques regroupent plusieurs millions d'articles (4 millions pour le serveur ADS[1]). La nécessité de tirer tout le potentiel scientifique de ces très grandes masses de données a conduit ces dernières années au développement du concept d'«observatoire virtuel astronomique». Celui-ci peut être défini comme *une entité destinée à rendre possible et à coordonner le développement des outils, des protocoles et des collaborations nécessaires pour réaliser tout le potentiel scientifique des données astronomiques dans la décennie à venir* (traduit du *National Virtual Observatory White Paper,* juin 2000). La mise en place d'un observatoire virtuel astronomique se heurte à différents verrous technologiques : la gestion de l'hétérogénéité des données (images, spectres et séries temporelles contenus dans les archives d'observatoires, résultats publiés dans les journaux électroniques, les bases de données de compilation, …), le développement de méthodes de réduction de données hyperspectrales, provenant en particulier des spectrographes à intégrale de champ, en tenant compte des spécificités des données astronomiques (grande dynamique des observations, objets intrinsèquement multi-échelle, caractéristiques du capteur connues,…), …

Le projet Masses de Données en Astronomie[2], soutenu par l'action concertée Incitative Masses de Données[3] du Ministère de la Recherche et de la Technologie, 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. Il repose sur la collaboration entre laboratoires de différents domaines : astronomie, recherche d'information, gestion des connaissances, réduction d'images hyperspectrales, calculs parallèles distribués. Le *Centre de Données astronomiques de Strasbourg* (CDS)[4], qui développe des services de référence largement utilisés par la communauté scientifique internationale, est au centre du projet.

Un travail considérable a déjà été accompli par la communauté pour définir des standards d'échange, des méta-données et des bases de connaissances terminologiques de type thésaurus. Un des objectifs principaux du projet, en lien avec nos travaux, est l'utilisation de ces ressources existantes pour l'élaboration de ressources de connaissance permettant d'indexer les données hétérogènes du domaine. La mise en place de cet objectif passe par plusieurs étapes. Elle nécessite la validation de deux bases de connaissances créées pour des applications différentes : les *Unified Content Descriptors* UCD (méta-données utilisées pour définir les colonnes des

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

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

[3] http://acimd.labri.fr/

[4] http://cdsweb.u-strasbg.fr/

catalogues) et le *thésaurus IAU* (utilisé pour l'exploration et l'indexation de la littérature publiée par référence à ces bases de connaissances), la réflexion sur la construction d'une ontologie du domaine et la prise en compte de l'intégration d'un domaine interdisciplinaire (les données de physique atomique et moléculaire d'intérêt astrophysique).

1.1.2 Participation de l'équipe EVI-SIG IRIT

Par sa participation au projet, l'équipe EVI-SIG propose de mettre à profit son expertise dans le domaine de la Recherche d'Information. Cette expertise est utilisée dans le développement de méthodes d'élaboration d'ontologies à partir des techniques de traitement des textes issues du domaine. L'utilisation de ces ressources a comme objectif d'aider les documentalistes dans leur tâche d'indexation par la proposition de mots clés, les documentalistes utilisant jusqu'ici la consultation manuelle du thésaurus de l'astronomie IAU. Afin de définir un langage d'indexation pouvant s'appliquer aux données hétérogènes, interprétable à la fois par les humains et les systèmes, nous proposons l'utilisation d'ontologies. Les ontologies représentent un consensus sur la connaissance du domaine. La contribution de l'équipe SIG-EVI consiste à envisager une méthode d'élaboration d'une ontologie pour définir un langage d'indexation du domaine au niveau conceptuel, permettant une indexation et une restitution des documents efficaces. Cette méthode repose sur la transformation et la mise à jour du thésaurus IAU, créé en 1995. Elle repose sur l'extraction d'informations à partir de données bibliographiques du domaine. Les mécanismes développés à la suite de ce travail seront ensuite étendus pour être appliqués à d'autres données que les données bibliographiques, notamment en les utilisant pour l'indexation des catalogues d'objets à partir de l'intégration des UCD dans le système.

1.1.3 Thème et tâche

Dans le cadre du projet MDA, nous avons spécifié les besoins auxquels doit répondre l'utilisation d'ontologies. Ces ressources doivent, d'une part, aider à interpréter le contenu des granules documentaires hétérogènes à partir de la connaissance du domaine et, d'autre part, aider l'utilisateur à cibler les données l'intéressant dans le large potentiel d'informations à sa disposition. Plus particulièrement, elles visent à spécifier un langage d'indexation permettant d'améliorer l'indexation et l'accès aux données disponibles dans le domaine et à focaliser la recherche sur les données intéressant particulièrement l'utilisateur dans la tâche de recherche qu'il entreprend.

Nous proposons d'appliquer, dans le cadre du projet, la modélisation du contexte d'une recherche présentée dans le chapitre 4. Cette modélisation repose sur une représentation de la connaissance associée au thème traité dans le corpus et sur une représentation de la connaissance associée à la tâche de recherche effectuée par l'utilisateur. Le domaine du thème abordé dans le corpus est relatif à la connaissance liée à l'astronomie abordée dans le corpus. La tâche sur laquelle nous nous concentrons est l'activité de veille sur les publications scientifiques. Elle est réalisée par des utilisateurs souhaitant analyser les activités de recherche dans le domaine ainsi que leurs acteurs et leur évolution [Hernandez 2004c].

1.2 Ressources existantes

Comme nous l'avons évoqué, des ressources ont été élaborées dans le domaine de l'astronomie et nous les décrivons plus en détail dans cette section.

1.2.1 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.

Le thésaurus est composé de termes considérés comme faisant partie du vocabulaire courant des astronomes. Ces termes ont d'abord été extraits de textes de référence, puis modifiés et mis à jour par des experts à partir de la première validation du thésaurus en 1992. Le thésaurus inclut des termes ayant directement un lien avec l'astronomie. Les termes couvrant des aspects limitrophes du domaine tels que l'histoire, l'instrumentation, l'électronique, l'informatique, la physique et les recherches spatiales, ou ayant un rapport avec les fusées ne sont pas intégrés car ils sont présents dans d'autres thésaurus tels que le Thésaurus de la NASA (NASA SP- 7064) et le INSPEC Thésaurus (1987). La norme utilisée pour concevoir le thésaurus est la norme ANSI Z39. Le thésaurus est disponible sous format texte et sous format HTML[5].

Un extrait du thésaurus IAU est présenté dans la figure 7.1

A DWARF STARS
BT A STARS
DWARF STARS
RT A SUBDWARF STARS

A STARS
BT EARLY TYPE STARS
NT A DWARF STARS
A GIANT STARS
A SUBDWARF STARS
A SUBGIANT STARS
A SUPERGIANT STARS
Ae STARS
Am STARS
Ap STARS
RT EARLY TYPE VARIABLE STARS
SPECTRAL TYPES

Légende
BT : terme plus spécifique
NT : terme plus générique
RT : terme lié

Figure 7.1 Extrait du thésaurus IAU

Ce thésaurus présente plusieurs inconvénients. Il a été réalisé en 1995 et ne contient pas les termes qui sont apparus depuis lors dans le domaine. Les relations vagues *RT* sont, de plus, difficilement exploitables automatiquement car les relations sémantiques qu'elles sous-entendent ne sont pas précisées. Comme les thésaurus en général, il contient la terminologie du domaine et non pas une conceptualisation du domaine. Sa migration vers une ontologie légère de domaine permettrait, d'une part, de le représenter dans une formalisation à la fois interprétable par les systèmes et les humains et, d'autre part, de proposer des mécanismes d'indexation sémantique à partir de la connaissance qu'il représente.

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

1.2.2 ADS

Le serveur ADS[6] (Astrophysics Data System) est une librairie électronique financée par la NASA regroupant plus de 4,4 millions d'articles du domaine. Son principal intérêt est de donner l'accès en ligne à ces documents. Il intègre un outil de recherche d'information traditionnel permettant la formulation de requêtes par mots clés. Les articles sont restitués par le système à partir des termes de la requête recherchés dans le contenu des documents et de méta-données comme les noms des auteurs, l'année de publication et les instances d'objets astronomiques mentionnés dans le contenu des publications. Un lien vers la base de données d'objets SIMBAD[7] est disponible.

L'outil de recherche reste limité dans ses performances car il n'intègre pas l'utilisation d'ontologies permettant de désambiguïser les termes des requêtes.

1.2.3 Autre ressource

Une ontologie légère du domaine de l'astronomie, factguru, est disponible sur le web[8]. Elle a été créée par un amateur de l'astronomie. Elle est constituée de concepts définis à partir d'un seul label et de relations taxonomiques et associatives entre les concepts. Sa composante lexicale étant très faible, son utilisation à des fins de RI est limitée. De plus, la connaissance qu'elle représente n'a pas été validée par des astronomes. Uniquement disponible à partir de pages HTML, elle n'est pas directement utilisable pour des traitements automatiques.

1.3 Evaluations

Les méthodes proposées dans cette thèse s'appliquant aux problématiques du projet MDA ont été évaluées. Cette section présente les éléments que nous avons validés dans ce contexte.

1.3.1 Eléments évalués en regard de nos propositions

L'évaluation des propositions présentées dans les chapitres 4, 5 et 6 consiste à valider les résultats obtenus par des experts du domaine. Afin de faciliter leurs évaluations, les éléments clés de ces propositions sont analysés. Ils sont évalués sur des échantillons de données. Les conclusions tirées permettent de les valider et d'étendre leur application à l'ensemble des données du domaine traité.

Le choix et la représentativité des éléments évalués pour chacune de nos propositions sont présentés dans les sections suivantes.

1.3.1.1 Modélisation du contexte à partir d'ontologie

La modélisation du contexte d'une recherche présentée dans le chapitre 4 repose sur l'utilisation de deux ontologies de domaine : l'une d'elle est dédiée à la tâche pour laquelle la recherche est effectuée, l'autre représente la connaissance du thème traité dans le corpus. L'ontologie proposée pour représenter la tâche de veille traitée dans le cadre du projet a été construite en collaboration avec des astronomes à partir des travaux présentés dans [Mothe 2004]. Sa pertinence a donc été validée lors de sa création. L'ontologie du thème utilisée dans le contexte de l'astronomie est élaborée à partir du thésaurus par les techniques proposées dans le chapitre 5. Les éléments de sa validation sont présentés dans la section 1.3.1.2.

---

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

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

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

L'intégration du modèle dans le système est réalisée par l'indexation des granules à partir du contexte spécifié par les deux ontologies. L'originalité de l'indexation proposée est l'utilisation des concepts de l'ontologie de thème comme descripteurs des granules. Afin de valider le mécanisme proposé, le résultat de cette indexation obtenue pour 10 documents du domaine est analysé. Ces documents sont choisis pour avoir été indexés manuellement par un même mot clé. Leurs thématiques sont donc a priori liées. L'ontologie utilisée pour réaliser l'indexation est l'ontologie obtenue à partir de la transformation du thésaurus. Afin que l'évaluation soit représentative d'un consensus dans le domaine, elle est réalisée par deux astronomes. Le protocole utilisé pour réaliser cette évaluation et les conclusions qu'elle permet de tirer sont présentés dans la section 4.

L'accès à l'information élaborée à partir des ontologies modélisant le contexte est validé par la création d'un prototype. Ce prototype est présenté dans le chapitre 8.

1.3.1.2 Transformation d'un thésaurus en une ontologie

La méthodologie permettant la transformation d'un thésaurus en une ontologie légère, présentée dans le chapitre 5, repose sur la définition de plusieurs règles. La validation de ces règles est réalisée par leur application à la transformation du thésaurus IAU. Les résultats obtenus pour chacune d'entre elles sont évalués par deux astronomes. Le protocole d'évaluation ainsi que les résultats obtenus pour les différentes règles sont présentés dans la section 2.

1.3.1.3 Adéquation entre un corpus et une ontologie

La méthodologie ainsi que les méthodes permettant d'analyser l'adéquation entre un corpus et une ontologie, présentées dans le chapitre 6, ont été appliquées pour valider l'utilisation de l'ontologie légère du domaine de l'astronomie pour l'indexation de deux corpus de publications du domaine. Elles sont validées par la réalisation d'un prototype. Ce prototype est présenté dans le chapitre 8. Nous avons utilisé la méthodologie et les méthodes pour comparer la connaissance représentée dans l'ontologie créée à partir du thésaurus et la connaissance représentée dans l'ontologie factguru. L'adéquation de ces deux ontologies a été évaluée par rapport à un corpus contenant des articles relatifs aux Rayons X. Les détails de cette analyse ne sont pas fournis dans ce chapitre. Ils ont montré des lacunes dans la connaissance représentée dans l'ontologie factguru quant à son adéquation au corpus (mauvaise adéquation lexicale et conceptuelle) et ont confirmé le besoin d'élaborer une nouvelle ontologie légère.

L'originalité des méthodes d'analyse de l'adéquation porte sur la pondération des concepts par leur représentativité sémantique dans un granule documentaire. Cette pondération repose sur la définition d'une mesure évaluant la proximité sémantique entre concepts dans l'ontologie par rapport au corpus. La mesure de proximité est analysée dans la section 3 de ce chapitre qui décrit le protocole d'évaluation et les résultats obtenus.

1.3.2 Corpus d'évaluation

La méthodologie de transformation d'un thésaurus en ontologie légère de domaine repose sur l'utilisation de corpus de référence. Deux corpus du domaine de l'astronomie ont alors 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. Les deux corpus ont été validés par des experts du domaine pour décrire les connaissances à représenter dans l'ontologie.

Comme nous l'avons évoqué dans les chapitres précédents, nos propositions reposent sur l'analyse syntaxique des corpus. L'analyseur syntaxique utilisé s'appuie sur un apprentissage endogène lui permettant de s'adapter à des corpus de différents domaines. Cependant, la spécificité des termes du domaine de l'astronomie a nécessité un pré-traitement des documents. Par exemple, dans le syntagme « *a galaxy* », la particule « *a* » ne représente pas un article indéfini mais la caractérisation d'une galaxie. Le pré-traitement a donc consisté à identifier, en collaboration avec des astronomes, les termes posant problème et à les spécifier en entrée à l'analyseur.

2 Transformation du thésaurus IAU en ontologie

Nous présentons dans cette section un retour d'expérience sur l'application de la méthode de transformation d'un thésaurus en ontologie légère présentée dans le chapitre 5. La transformation du thésaurus IAU a été réalisée à partir de l'ensemble des règles de transformation que nous avons définies.

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

2.1 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 les 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. Le format utilisé appliqué à un exemple d'évaluation est présenté dans la figure 7.2. Les résultats sont ensuite dépouillés à partir des fichiers annotés par les experts.

Objectif : valider les nouvelles relations décelées entre concepts descendant du type abstrait « property »

Pour le concept density, nouvelles relations avec
    le concept mass
    le concept perturbation
    …

Pour le concept spectra, nouvelles relations avec
    le concept quasar
    le concept ccd
…

Figure 7.2 Exemple de fichier d'évaluation des règles de transformation du thésaurus

2.2 Concepts extraits du thésaurus

Les règles R1, R2 et R3 permettent la création de concepts à partir des termes liés par des relations de synonymie (relations « *Utiliser plutôt* » et *« Utiliser pour désigner »)* dans le thésaurus.

En s'appuyant sut les relations de ce type définies sur les 2957 termes du thésaurus, 2547 concepts ont été crées. Ces concepts ont entre 1 et 4 labels.

La pertinence des concepts créés et définis à partir de plusieurs labels a été validée par les astronomes. La validation a montré que la totalité des concepts créés étaient pertinents et que, pour 85 % d'entre eux, l'ensemble des labels était correct. Pour 15 %, les labels ne sont pas

corrects car ils se rapportent à des sous-concepts des concepts pour lesquels ils sont définis. Ces labels ont donc été supprimés et ont mené à la création de nouveaux concepts définis comme sous-concepts des concepts auxquels ils étaient rattachés dans la première proposition.

2.3 Hiérarchie de concepts

L'organisation hiérarchique des concepts est réalisée par les règles R4 et R5. Les relations sont définies à partir de la relation « est plus spécifique », « est plus générique » du thésaurus.

Ces relations ont mené à la définition de 2882 relations « sous-classe de » dans l'ontologie. Parmi celles-ci, 193 relations redondantes ont été trouvées. Elles ont donc été supprimées de la hiérarchie de concepts. Les relations étant définies dans les spécifications du thésaurus pour ne comprendre que des relations du type « est plus spécifique », « est plus générique », seules 5% de ces relations ont été analysées. La totalité a été jugée valide.

2.4 Types abstraits

Au niveau le plus générique de l'ontologie créée à ce stade de la transformation, 1132 concepts ont été définis. Ce nombre élevé rend difficile la navigation dans la hiérarchie de l'ontologie dans la mesure où le nombre de concepts au premier niveau de la hiérarchie est trop important. De nouveaux concepts d'un niveau plus général sont alors ajoutés. Ces concepts sont obtenus par deux mécanismes appliqués successivement. Les résultats obtenus sont présentés dans les sections suivantes.

2.4.1 Regroupements des concepts de haut niveau à partir de la tête de leur label

Les règles R6 et R7 permettent de créer un nouveau niveau hiérarchique à partir de la tête commune des labels des concepts de haut niveau dans l'ontologie. Cette transformation mène à la création d'un nouveau niveau hiérarchique composé de 682 concepts. Ce nombre restant important, le mécanisme de création de types abstraits présenté dans la section suivante est alors mis en place.

2.4.2 Types abstraits proposés

La règle R8 et R9 permettent d'obtenir les types abstraits des concepts du plus haut niveau de généralisation dans l'ontologie 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. Parmi ces types, 14 ont été retenus. Ils sont présentés dans la figure 7.3. Notons que seule la définition du type *instrumentation* a été modifiée car elle introduisait à l'origine une notion de localisation fausse pour l'astronome.

Property : a basic or essential attribute shared by all members of a class

Phenomenon : any state or process known through the senses rather than by intuition or reasoning

Event : *something that happens at a given time*

Science : a particular branch of scientific knowledge

Instrumentation : an artifact (or system of artifacts) that is instrumental in accomplishing some end

Substance : that which has mass and occupies space

Relation : an abstraction belonging to or characteristic of two entities or parts together

Location : a point or extent in space

Angle : the space between two lines or planes that intersect; the inclination of one line to another

Plane : an unbounded two-dimensional shape

Region : the extended spatial location of something;

Object : a tangible and visible entity

Natural object : an object occurring naturally; not made by man

Artefact : a man-made object taken as a whole

Figure 7.3 Types abstraits extraits de WordNet

Les concepts abstraits ainsi extraits ont été comparés aux concepts abstraits définis dans l'ontologie existante factguru. Les astronomes ont validé la pertinence des concepts que nous avons proposés. Comparés à ceux de l'ontologie existante, ils sont beaucoup plus précis et représentatifs du domaine.

La figure 7.4 présente l'organisation hiérarchique des types également issus de WordNet et validée par les astronomes.

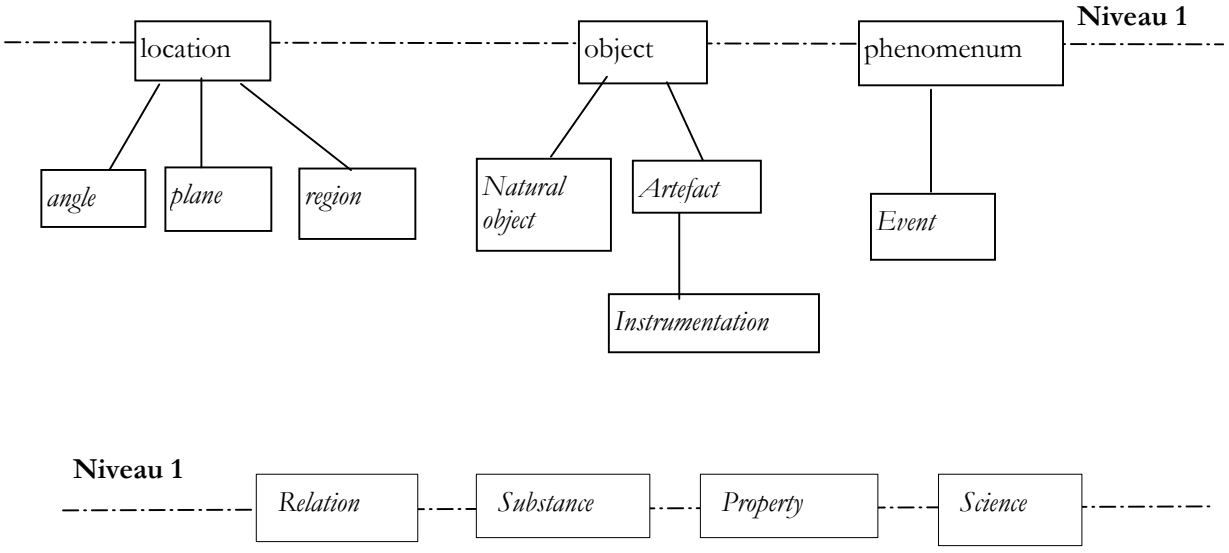


Figure 7.4 Organisation hiérarchique des types

Des échantillons des concepts rattachés à chacun des types en italique dans la figure 7.4 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 7.1

Types abstraits de rattachement	**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%

Tableau 7.1 : 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 ne sont pas d'accord avec les hyperonymes associés par WordNet aux synsets définis à partir des labels de ces concepts de l'ontologie descendant de ce type abstrait.

Dans le cas des concepts qui ne sont pas correctement attachés à un type abstrait (soit 11% des concepts évalués), les astronomes proposent un nouveau type dans l'ensemble des types abstraits proposés. Les concepts sont alors manuellement rattachés au type proposé par l'expert. Si le concept ne peut être rattaché à aucun des types, il est associé à un type abstrait « sans type » indiquant que son type devra être ré-étudié par la suite. En considérant que l'ensemble des concepts de l'ontologie hérite du type abstrait de leur ancêtre le plus générique, le procédé que nous avons mis en place permet d'associer un type abstrait à 50% des concepts définis dans l'ontologie. Ce résultat est intéressant car il est possible d'élaborer une première base pour l'ontologie pour laquelle les types sont définis. Un mécanisme devra par la suite être développé soit pour définir de nouveaux types abstraits, soit pour proposer un mécanisme d'association aux types abstraits pour les concepts dont les labels n'apparaissent pas dans WordNet. Les règles de détection des types abstraits et de leur association aux concepts de l'ontologie permettent néanmoins d'aider considérablement dans la création de l'ontologie car les types abstraits sont proposés automatiquement ainsi que leur rattachement aux concepts.

2.5 Spécification des relations associatives entre concepts

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

2.5.1 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 coté desquels ils apparaissent. L'ensemble du contexte associé aux types *property* et *phenomenon* est présenté dans le tableau 7.2. 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.

Contexte et nature des termes	**Property**	**Instrumentation**
Verbes dont les labels sont objets	Have Measured by Estimate Increase,decrease influences	Observed with Observed by Perform Carry out with By means of
Syntagmes nominaux	Variation Distribution Determination Source Measurement	Observation

Tableau 7.2 : Termes apparaissant fréquemment dans le contexte des concepts rattachés aux types abstraits property et phenomenon dans le corpus documentaire

La spécification des relations entre types abstraits a nécessité deux heures de travail pour les experts. Les relations définies pour les types *property* et *phenomenon* et les autres types abstraits sont présentés dans le tableau 7.3.

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

	Substance	**Relation**	**Angle**	**Plane**	**Region**	**Artefact**
Property	*Is a property of*	*Is a property of*	*Is a property of*	*Is a property of*	*Is a property of*	*Is a property of*
Instrumentation	*Observes Measures*	*Observes Measures*	*Observes Measures*	*Observes Measures*	*Observes Measures*	*Has or is part exclude*

Tableau 7.3 Relations pour les types abstraits *property* et *instrument*

2.5.2 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 7.4.

	Nombre de relations vagues évaluées	**Nombre de relations qui ne sont pas correctement labellisées**
Concepts descendant du type abstrait property	34	5
Concepts descendant du type abstrait property	15	3

Tableau 7.4 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.

2.5.3 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 de 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*. Le tableau 7.5 récapitule le résultat de l'évaluation des relations que nous proposons.

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

Tableau 7.5 Résultat de l'analyse des nouvelles relations entre concepts proposées à partir du contexte de leur label dans le corpus

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

La deuxième méthode d'extraction du contexte d'un label repose sur l'analyse distributionnelle réalisée par le module UPERY de Syntex. Ce type d'analyse consiste à rapprocher des labels en fonction de la ressemblance de leur contexte. Ils sont rapprochés s'ils sont formés dans le corpus autour des mêmes relations syntaxiques et des mêmes têtes et queues. Un coefficient de proximité entre termes est défini dans le module. Ce coefficient tient compte de la productivité des contextes partagés par les termes, la productivité correspondant aux nombres de termes qui partagent le contexte. Le coefficient de proximité repose sur le principe suivant : si un contexte partagé par deux termes est très productif, sa contribution au rapprochement des deux termes est a priori plus faible que celle d'un contexte peu productif. Nous avons donc sélectionné, pour évaluer notre approche, les termes rapprochés par un coefficient de proximité inférieur ou égal à 80% de la proximité maximale.

Le tableau 7.6 présente les résultats de l'évaluation des relations ainsi décelées pour les concepts descendants des types *property*.

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

Tableau 7.6 Résultat de l'analyse des nouvelles relations entre concepts proposées à partir du contexte de leur label dans le corpus

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

2.6 TPertinence 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.

2.6.1 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.

2.6.2 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 7.7. Les résultats de l'intégration des nouveaux termes en tant que relations associatives entre concepts existants sont présentés dans le tableau 7.8.

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

Tableau 7.7 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%

Tableau 7.8 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.

2.7 Bilan

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. Les méthodes de mise à jour permettent également d'aider à la détection de nouveaux termes pertinents et à la proposition de leur intégration dans l'ontologie.

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. Finalement, la mesure de sélection des nouveaux termes spécifiques au corpus devra être améliorée pour entraîner moins de bruit.

3 Mesure de proximité entre concepts dans une ontologie

La mesure proposée pour pondérer la proximité sémantique entre deux concepts d'une ontologie a été présentée dans le chapitre 4 section 3.1.1.3. Son principal intérêt est de prendre en compte l'ensemble des relations entre concepts définis dans l'ontologie (relations associatives et relations taxonomiques) et de pondérer les liens entre concepts à partir de l'information sur les concepts contenue dans le corpus.

Le protocole défini pour évaluer cette mesure est présenté dans la section 3.1. Les résultats obtenus par cette mesure sont présentés dans la section 3.2.

3.1 Protocole d'évaluation

Le protocole mis en place pour évaluer la mesure de proximité entre concepts proposées dans le chapitre 4 consiste à comparer les résultats obtenus par la mesure à des jugements utilisateurs. Cette démarche est employée dans de nombreuses études [Resnik 1995], [Jiang 1997], [Lin 1998], [Budanitsky 2001]. Cependant, comme nous l'avons souligné dans le chapitre 3, les évaluations de ces études consistent à comparer la proximité entre paires de termes et non pas entre paires de concepts. L'interprétation de la proximité sémantique est alors biaisée par le choix subjectif de l'utilisateur concernant le concept auquel se rapporte un terme. Pour y remédier, le protocole d'évaluation que nous avons choisi consiste à faire évaluer, par les utilisateurs, la proximité entre paires de concepts. Afin que les concepts soient interprétés dans le sens où ils sont définis dans l'ontologie, les concepts sont présentés aux utilisateurs avec leur contexte local. Ce contexte est composé de l'ensemble des labels du concept et de l'ensemble de ses relations directes (labels du concept, concepts pères, concepts fils, et concepts auxquels ils sont liés par des relations associatives).

Le protocole d'évaluation consiste à faire évaluer à trois astronomes la proximité sémantique entre 20 paires de concepts dont le contexte est spécifié. Un extrait du fichier donné à évaluer aux astronomes est présenté dans la figure 7.5.

Paires de concepts		**Proximité sémantique Valeur entre 0 et 4**
Concept 1	**Concept 2**	
Concept = X_ray *Labels* : X ray, X ray radiation *Concepts pères* : electromagnetic_waves, high_energy_radiation,	**Concept = waves** *Labels* : wave *Concepts fils* :gravity_waves, drift_waves, electromagnetic_waves, submillimeter_waves, acoustic_waves, gravitational_waves, shock_waves, plasma_waves, electrostatic_waves, ion_plasma_waves,	
Concept = X_ray *Labels* : X ray, X ray radiation *Concepts pères* : electromagnetic_waves, high_energy_radiation	**Concept = electromagnetic_waves** *Labels* : electromagnetic wave, *Concepts pères* :waves, Concepts fils :radio_waves, ultraviolet_radiation, X_rays, cerenkov_radiation, millimeter_waves, light, bremsstrahlung, gamma_rays,	
Concept = X_ray *Labels* : X ray, X ray radiation *Concepts pères* : electromagnetic_waves, high_energy_radiation	**Concept = photons** *Labels* : photon, *Concepts pères* :bosons,	

Figure 7.5 Extrait de la fiche d'évaluation portant sur la proximité sémantique entre concepts

Les valeurs de proximité données par les astronomes sont ensuite comparées aux valeurs obtenues par la mesure proposée. La mesure est appliquée à l'ontologie de l'astronomie que nous avons obtenue par la transformation du thésaurus. Pour cette évaluation, l'ontologie n'a pas été mise à jour car les propositions de cette étape n'ont pas été validées entièrement par les astronomes. Les résultats de la comparaison sont présentés dans la section 3.2.

3.2 Comparaison aux jugements humains

L'évaluation présentée dans cette section vise à évaluer si la proximité affectée par la mesure est cohérente avec les jugements de proximité des astronomes. La mesure proposée prend en compte l'information contenue sur les concepts dans le corpus considéré. Afin d'isoler ce facteur, les valeurs considérées pour la mesure, dans cette partie de l'évaluation, correspondent aux résultats obtenus à partir du corpus contenant des résumés d'articles publiés en 1995. La période de publication des résumés est commune à celle de l'élaboration du thésaurus. L'ontologie considérée dans l'évaluation n'ayant pas été mise à jour, la connaissance qu'elle représente se rapporte à cette époque.

Les valeurs de proximité sémantique moyennes affectées aux paires de concepts par les astronomes sont corrélées aux valeurs affectées par la mesure proposée.

La mesure de corrélation utilisée (1) est la mesure considérée dans [Lord 2003].

$$corr(prox\_mes, prox\_astro) = \frac{\sum (prox\_mes_i - \overline{prox\_mes})(prox\_astro_i - \overline{prox\_astro})}{\sqrt{\sum (prox\_mes_i - \overline{prox\_mes})^2 \sum (prox\_astro_i - \overline{prox\_astro})^2}} \quad \textbf{(1)}$$

où $prox\_mes_i$ correspond aux résultats obtenus pour les différentes paires *i* par la mesure de proximité, $prox\_astro_i$ aux résultats obtenus pour les différentes paires *i* par les astronomes et $\overline{prox\_mes}$ (resp. $\overline{prox\_astro}$) correspond à la moyenne des mesures obtenues par la mesure (resp. par les astronomes).

Nous avons corrélé les valeurs de proximité obtenues par la mesure et les valeurs fournies par deux astronomes. Les résultats sont présentés dans le tableau 7.9. Afin de déterminer si la mesure proposée se rapporte à un consensus sur la proximité entre concepts dans le domaine, nous avons considéré la moyenne des valeurs affectées par les deux astronomes. Ces valeurs correspondent à l'astronome « virtuel » dans le tableau de résultats. La mesure que nous proposons est également comparée à deux mesures de la littérature : la mesure originalement proposée par Jiang [Jiang 1997] qui ne prend en compte que les relations taxonomiques et la mesure proposée par Lord [Lord 2003] qui considère avec le même poids les relations taxonomiques et non taxonomiques.

Corrélation	Astronome « virtuel »
Mesure proposée	0,47
Mesure de Jiang [Jiang 1997] ne prenant en compte que les relations « est un »	0,11
Mesure de Lord[Lord 2003] considérant toutes les relations comme des relations « est un »	0,32

Tableau 7.9 Résultats de la comparaison des valeurs de proximité affectées par des humains et celles affectées par la mesure proposée et les mesures de la littérature

La mesure que nous proposons est mieux corrélée aux valeurs données par les astronomes que les mesures de la littérature. Ceci s'explique notamment par la validation de deux hypothèses qui nous ont mené à la spécification de cette mesure. Tout d'abord, les relations non taxonomiques doivent être prises en compte dans le calcul de la proximité sémantique. Par exemple, dans l'ensemble des concepts que nous avons considéré, les astronomes affectent une

proximité de 2,5 aux concepts *X ray* et *X ray absorption* Ces concepts sont reliés par la relation non-taxonomique « is a phenomenon linked to » et sont subsumés par le concept générique « phenomenon ». Dans le cas où les relations non taxonomiques ne sont pas prises en compte, la proximité calculée entre ces deux concepts est très faible car le concept qui les subsume est très générique. Par contre, par notre mesure, la proximité calculée est plus importante car la relation non taxonomique permet d'établir un lien sémantique direct entre les deux concepts. Ensuite, l'intuition selon laquelle deux concepts liés par plusieurs relations taxonomiques n'ont pas forcément de l'information en commun est mise en valeur par les résultats de la corrélation obtenue par la mesure de Lord. En prenant, par exemple, les concepts *X ray* et *radio source* qui sont certes liés par deux relations non taxonomiques dans l'ontologie, les astronomes affectent comme proximité 0 entre ces deux concepts (ce que fait également notre mesure). De plus, la mesure de Lord affecte des valeurs de proximité de manière générale plus faible et moins discriminante entre les concepts car elle repose sur le contenu en information des concepts calculés à partir de tous les concepts auxquels ils sont liés.

Notons cependant que les coefficients de corrélation obtenus sont plus bas que ceux obtenus dans les travaux de la littérature qui sont généralement autour de 0,8 [Lin 1998] [Jiang 1997] et [Mc Hale 1998]. Ceci s'explique en partie par le fait que l'ontologie utilisée n'a pas été mise à jour avec de nouvelles relations et termes détectés dans le corpus. L'ajout de ces éléments permettra d'améliorer la qualité des valeurs restituées par la mesure. Une autre explication est que le coefficient de corrélation entre les valeurs affectées par les deux astronomes est de 0,54 également plus bas que les coefficients calculés entre les jugements humains utilisés pour réaliser les évaluations de la littérature [Resnik 1995]. De nouvelles expérimentations devront être réalisées à partir de l'ontologie mise à jour et du jugement d'autres experts du domaine.

4 Indexation sémantique des documents suivant la modélisation du contexte

Le procédé d'indexation sémantique présenté dans le chapitre 4 consiste à extraire les concepts des documents et à les pondérer à partir de la mesure de représentativité proposée. Cette mesure prend en compte à la fois la représentativité statistique et la représentativité sémantique du concept dans le granule. L'indexation sémantique est réalisée à partir de l'ontologie légère résultant de la transformation du thésaurus sans intégrer les mises à jour car ces modifications n'ont pas encore été toutes validées par les astronomes.

4.1 Protocole

Le protocole est composé de deux étapes.

La première étape consiste à considérer un ensemble de documents et à proposer pour chacun d'entre eux un ensemble de dix concepts ayant le plus fort score de représentativité résultant de l'indexation. L'expert doit alors valider la pertinence des concepts proposés pour chacun des documents, puis classer les concepts par leur degré de pertinence. Un extrait du fichier donné à évaluer aux astronomes est présenté dans la figure 7.6.

La seconde étape considère l'ensemble des documents et propose un ensemble de concepts représentatifs de l'ensemble. Cette étape vise à évaluer la représentativité d'un concept dans un granule composé de plusieurs documents. De la même façon que pour le granule document, l'expert doit valider la pertinence des concepts et les classer selon leur degré de pertinence.

L'accès aux contextes associés à chacun des concepts retenus pour indexer un granule documentaire aurait dû être présenté par le biais de l'interface aux astronomes. Cependant, des contraintes de proximité géographique n'ont pas permis de mettre en place ce type d'évaluation. Ces évaluations sont donc une première étape.

Document	Concepts par ordre alphabétique
DOC-1	accretion disks coude spectrograph detector event intensity neutron star observatory observation X ray X ray pulsar
DOC-2	group luminosity modulation microwave nonthermal source observatory radio source rotation star X ray

Figure 7.6 Extrait de l'évaluation de l'indexation sémantique proposée aux astronomes

4.2 Pertinence des concepts indexés pour un granule correspondant à un document

Pour chacun des documents, la pertinence des concepts proposés est analysée. Le tableau 7.7 présente les résultats obtenus.

Documents	Taux de précision[9]
DOC1	0,7
DOC2	0,8
DOC3	0,7
DOC4	0,5
DOC5	0,9
DOC6	0,7
DOC7	0,8
DOC8	0,8
DOC9	1
DOC10	0,8

Tableau 7.10 Résultats de l'indexation sémantique pour les différents documents

L'évaluation montre que les concepts sont correctement associés à chaque document avec une précision moyenne de 0,8. Afin d'évaluer la pertinence de la mesure de représentativité d'un concept dans un document, nous avons également analysé si l'ordre dans lequel elle permet d'ordonner les concepts correspond effectivement à l'importance que l'expert accorde au concept dans le document. La comparaison des deux listes ordonnées met en valeur l'intérêt de la mesure.

[9] $\text{Précision} = \frac{\text{nombre de concepts correctements restitués}}{\text{nombre de concepts restitués}}$

En effet, pour la moitié des documents l'ordre des concepts correspond à un ou deux concepts près.

Bien que ces résultats soient encourageants, d'autres étapes dans la validation de l'indexation sont nécessaires. Par exemple, nous analyserons les concepts non identifiés dans les documents que l'expert jugerait pertinents ainsi que le seuil à fixer sur le score de représentatitivité à prendre en compte pour la restitution des concepts.

4.3 Pertinence des concepts indexés pour un granule correspondant à un ensemble de documents

Pour l'ensemble des documents considérés comme formant un seul granule, la pertinence des concepts est analysée.

L'évaluation a consisté, dans un premier temps, à analyser la pertinence des dix concepts restitués pour ce granule. L'expert a rejeté seulement un des dix concepts proposés. Ceci montre que l'indexation sémantique permet d'extraire en grande partie les concepts d'un granule quand celui-ci est composé de plusieurs documents. Dans la seconde étape, nous avons analysé le score de représentativité associé à ces concepts et l'importance que l'expert leur accorde pour représenter l'ensemble des informations qu'il avait lues dans les dix documents. Les deux classements sont très proches. Les concepts sont en effet situés soit dans la même position, soit à une place d'intervalle.

4.4 Bilan

Les évaluations que nous avons réalisées sont une première étape vers la validation du procédé d'indexation sémantique que nous proposons. Nous avons mis en évidence l'intérêt du mécanisme proposé par la pertinence des concepts extraits de chaque granule (qu'il soit composé d'un document ou d'un ensemble de documents), et la pondération proposée pour déterminer leur importance dans le granule.

D'autres évaluations seront nécessaires. Nous envisageons de fixer empiriquement un seuil sur le score de représentativité des concepts à considérer pour chaque granule. Nous souhaitons également analyser si le procédé d'indexation extrait tous les concepts se rapportant aux documents dans l'ontologie. Ces deux paramètres ne pourront être analysés qu'à partir du moment où l'ontologie de l'astronomie sera mise à jour (par l'ajout de nouveaux termes et de nouvelles relations) afin qu'elle contienne toute la connaissance nécessaire à cette analyse.

5 Conclusion

Dans le domaine de l'astronomie, le thésaurus IAU construit en 1995 fait référence et les acteurs du domaine souhaiteraient l'utiliser plus largement. Le fait qu'il ne soit pas à jour et que sa connaissance ne soit pas formalisée sont des freins à son utilisation. Dans le cadre de notre participation au projet Masses de Données en Astronomie, nous avons appliqué à ce thésaurus la méthode de transformation de thésaurus en ontologie légère présentée dans le chapitre 5. L'évaluation réalisée par les astronomes a permis de valider notre méthode. Les mécanismes qu'elle propose permettent d'extraire avec pertinence les concepts du domaine à partir de leurs différents labels ainsi que d'identifier les relations sémantiques entre eux. De plus, les expérimentations ont mis en valeur le mécanisme de mise à jour de la connaissance à partir de textes du domaine. Notre contribution au projet a amené un regain d'intérêt pour l'utilisation de ressources du domaine. Elle se concrétisera par sa présentation dans le cadre du projet de l'observatoire virtuel international[10] et au congrès de l'« International Astronomical Union » (IAU)[11].

Les parties de nos contributions relatives à la mesure de proximité sémantique entre concepts et à l'indexation sémantique n'ont pu être évaluées que sommairement par les astronomes. Ces aspects ne font que partiellement partie du projet. Cependant, nous avons pu souligner leur intérêt. Sur le corpus que nous avons considéré, le mécanisme d'indexation sémantique permet d'extraire avec pertinence les concepts de chaque granule et la mesure de pondération proposée est proche des jugements de l'expert. D'autres expérimentations devront être réalisées afin d'évaluer si la totalité des concepts sont extraits pour chaque granule. L'indexation devra également être analysée sur un corpus de granules plus important. Bien que l'interface d'exploration de corpus à partir des ontologies n'ait pas été évaluée, elle a été présentée aux astronomes. Ces derniers ont apprécié les fonctionnalités de navigation qu'elle propose. Ils ont souligné l'intérêt des deux niveaux d'exploration (concepts et instances) et ont insisté sur l'importance de la spécification du contexte associé à chacun des concepts (affichage de labels et de l'ensemble des relations avec les autres concepts) pour mieux appréhender le rôle d'un concept dans la modélisation du contexte. D'autres évaluations seront nécessaires pour analyser en détail l'ensemble des mécanismes de navigation proposés.

[10] http://www.ivoa.net/

[11] http://www.astronomy2006.com/

Chapitre 8
Prototype : OntoExplo

1 Introduction

Le prototype OntoExplo a été réalisé afin de mettre en place un système de RI reposant sur la modélisation du contexte d'une recherche présentée dans le chapitre 4. Il repose sur une représentation de la connaissance associée à la tâche dans laquelle s'inscrit la recherche (ontologie de la tâche) et sur une représentation de la connaissance associée au thème abordé dans le corpus (ontologie du thème) [Hernandez 2005b]. OntoExplo vise tout d'abord à mettre en œuvre la méthodologie et les méthodes définies dans le chapitre 6 pour évaluer l'adéquation entre une ontologie de thème et un corpus, permettant ainsi de valider le choix de l'ontologie pour la modélisation du contexte associé au corpus. Il implante également les mécanismes d'indexation sémantique à partir de l'ontologie du thème et les mécanismes d'annotation des documents à partir de l'ontologie de tâche décrits dans le chapitre 4. A chacun de ces mécanismes sont associées des interfaces permettant à l'utilisateur l'accès aux informations.

Le prototype peut s'appuyer sur n'importe quelle ontologie formalisée en OWL. Il est illustré dans ce chapitre à partir d'exemples issus de notre cas d'étude, le projet Masse de Données en Astronomie, décrit dans le chapitre 7.

Ce chapitre est organisé de la façon suivante. La section 2 décrit l'architecture du prototype composée de trois couches. La section 3 présente la première couche permettant l'accès et la manipulation d'ontologies. La section 4 décrit la couche suivante relative à l'implantation de l'adéquation entre une ontologie de thème et un corpus. La section 5 présente la dernière couche permettant l'intégration du contexte dans le traitement des documents. L'interface est, quant à elle, présentée dans chacune des sections correspondantes.

2 Architecture

L'architecture du prototype se décompose en trois couches comme l'indique la figure 8.1.

La première couche gère l'accès et la manipulation des ontologies, qu'elles soient de thème ou de tâche. Au bas niveau, la gestion des ontologies s'effectue par l'utilisation d'une API java permettant de rendre manipulables par le système les éléments d'une ontologie OWL en termes de classes et d'objets. Cette couche comprend également une interface à partir de laquelle un utilisateur peut visualiser et naviguer dans les ontologies.

Au niveau de la deuxième couche s'effectue l'analyse de l'adéquation entre l'ontologie du thème et le corpus. L'ontologie choisie doit, en effet, refléter le contexte associé au corpus. Le choix entre plusieurs ontologies pourra s'appuyer sur les résultats fournis par cette couche. Cette couche comprend l'implantation des méthodes décrites dans le chapitre 6 et une interface permettant à un utilisateur de visualiser les résultats de l'analyse.

La troisième couche intègre le contexte représenté par les ontologies aux traitements des documents. A ce niveau, une ontologie de thème et une ontologie de tâche sont associées au corpus. Les documents sont annotés à partir de la connaissance contenue dans les deux ontologies par les techniques décrites dans la section 3 du chapitre 4. Le résultat de l'annotation est stocké dans une base de données afin d'optimiser leur accès au cours de leur exploitation. Une interface est également intégrée à cette couche. Par son intermédiaire, l'utilisateur accède aux annotations et explore ainsi la collection de données.

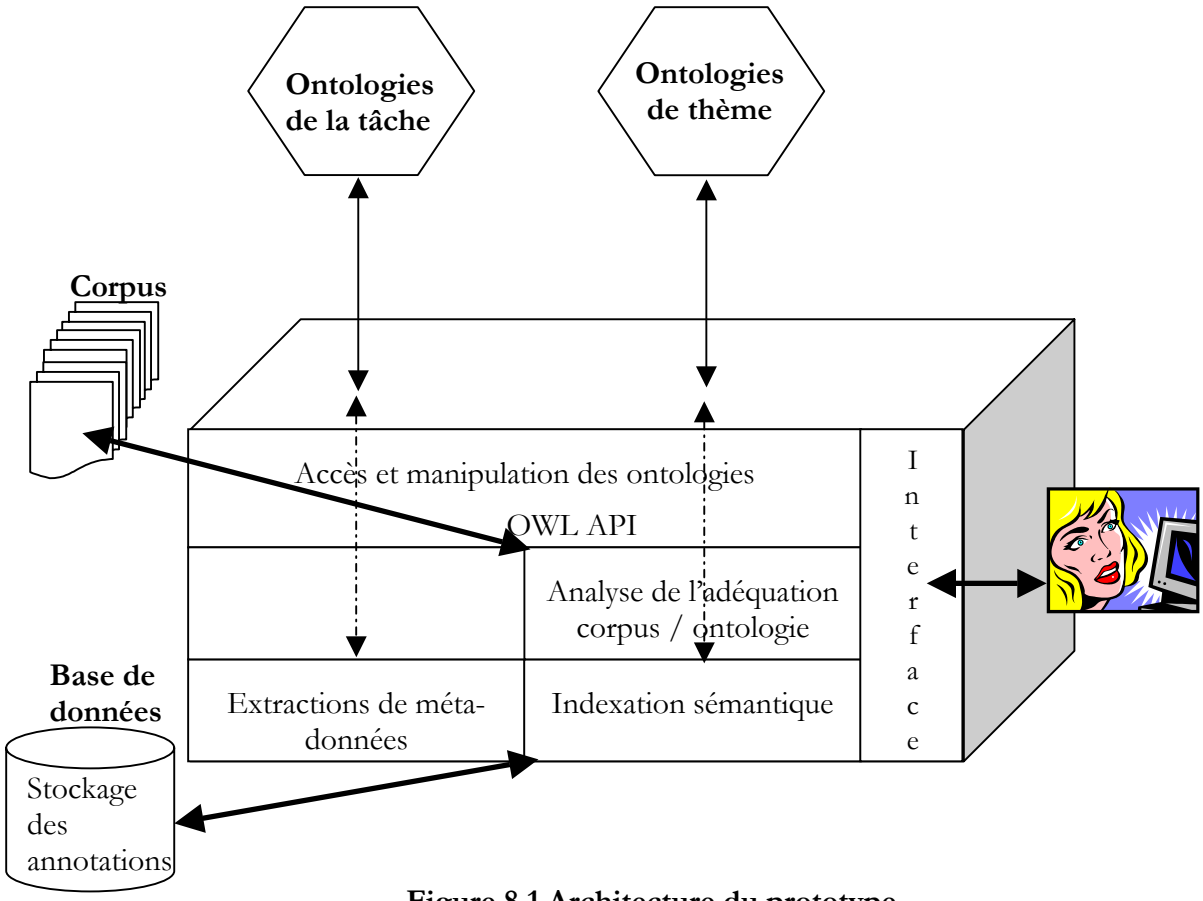


Figure 8.1 Architecture du prototype

Les différentes couches sont décrites en détail dans les sections suivantes.

3 Accès et manipulation des ontologies

La première couche du système vise, d'une part, à implanter l'accès et la manipulation des ontologies par le système et, d'autre part, à proposer une interface de visualisation et de navigation des ontologies par un utilisateur.

3.1 Implantation

L'implantation de l'accès et de la manipulation des ontologies au bas niveau doit permettre au système de gérer efficacement les éléments des ontologies légères que nous considérons (concepts, relations taxonomiques et associatives, instances). Ces éléments sont manipulés afin de permettre plusieurs traitements par le prototype : la visualisation et la navigation des ontologies par un utilisateur, le calcul de l'adéquation entre une ontologie et un corpus et l'annotation d'un corpus documentaire à partir d'ontologies.

La plate-forme d'édition Protégé 2000 [Noy 2000] est utilisée dans de nombreux travaux car elle permet l'édition d'ontologie OWL et l'ajout de modules appelés plug-ins permettant d'intégrer des traitements sur les ontologies (visualisation, inférence, navigation, …) [Wielinga 2001], [Thieu 2004]. Nous avons donc considéré l'utilisation de cette plate-forme. Cependant, le temps de chargement qu'elle implique pour une ontologie comportant plus d'un millier de concepts (environ une demi-heure, par exemple, pour l'ontologie de l'astronomie que nous avons élaborée grâce à notre méthodologie à partir du thésaurus IAU) la rend inexploitable pour les fins requises. Le temps de chargement important s'explique par les traitements nécessités par la plate-

forme pour la transcription de l'ontologie dans la représentation interne du système et pour son édition. L'étude présentée dans montre que les systèmes de connaissance reposant sur des représentations OWL, qu'ils stockent en mémoire l'ontologie ou de façon persistante dans une base de données, manquent d'efficacité quant au temps de chargement de l'ontologie dans le système ou au temps d'accès aux instances par des requêtes. A titre indicatif, une ontologie composée d'environ 100000 instances met plus de trois heures à se charger dans la mémoire virtuelle du système OWLJessKB utilisé sur un PC de capacité standard. Bien que les ontologies légères utilisées dans notre modèle nécessitent le stockage de moins de connaissance que les ontologies formelles utilisées dans [Guo 2004], il est indispensable que le système de stockage des annotations soit capable de gérer efficacement les concepts des deux ontologies du modèle et les instances correspondant aux annotations.

Nous nous sommes de plus attardés sur les plug-ins de visualisation d'ontologies associés à Protégé 2000 car les tâches de RI que nous considérons reposent sur la navigation au sein d'une ontologie en OWL.

Nous avons pour cela comparé les cinq plug-ins de visualisation suivant :

- ezOWL, réalisé par un groupe de l'Institut de Recherche en Electronique et Télécommunications de Corée[1],
- Jambalaya, conçu par le Laboratoire d'Ingénierie de Logiciels et d'Interactions Homme Machine (CHISEL) du Canada[2],
- OntoViz, développé par le Centre de Recherche Allemand d'Intelligence Artificielle[3],
- OWLViz, réalisé par un membre de l'Université de Manchester au Royaume-Uni[4],
- TGViz, conçu par un membre du Département d'Intelligence, d'Agents, de Multimédia, d'Electronique et de Science Informatique de l'Université de Southampton au Royaume-Uni[5].

La comparaison a porté sur l'affichage d'un élément, la possibilité de zoomer sur une partie de l'ontologie, l'affichage du graphe correspondant à l'ontologie, la gestion des préférences, les automatismes liés à l'affichage, la gestion des préférences. Le détail de l'évaluation se trouve dans [Cardoner 2004]. Deux caractéristiques générales sont ressorties :

- Il n'est pas possible de choisir les relations à afficher (toutes les relations sont affichées).
- On ne peut pas sélectionner un concept, ce qui empêche de pouvoir récupérer des évènements afin d'afficher de nouvelles informations en temps réel.

Nous avons donc choisi d'implanter un prototype indépendant d'une plate-forme existante afin de ne pas augmenter le temps de chargement d'une ontologie volumineuse par les traitements liés à la plate-forme et de proposer un mécanisme d'exploration qui répond aux limites que nous avons soulignées.

Nous avons utilisé l'API java OWL [Bechhofer 2003] qui est libre d'utilisation (open source) pour traiter les ontologies OWL. Cette interface de haut niveau permettant d'accéder et

---

[1] http://iweb.etri.re.kr/ezowl/

[2] http://www.thechiselgroup.org/~chisel/projects/jambalaya/jambalaya.html

[3] http://protege.stanford.edu/plugins/ontoviz/ontoviz.html

[4] http://www.co-ode.org/downloads/OWLVizPlugin.html

[5] http://www.ecs.soton.ac.uk/~ha/TGVizTab/TGVizTab.htm

de manipuler des ontologies OWL a été conçue pour être un composant réutilisable par toutes les applications intégrant ce type d'ontologies. Elle a été proposée afin de répondre à 5 principales attentes :

- l'analyse des documents OWL correspondant à la transformation d'une représentation concrète d'un document OWL en une représentation interne,
- la modélisation des éléments OWL fournissant des structures de données représentant ou encodant ces éléments en facilitant leur accès,
- la manipulation des éléments OWL fournissant des représentations et des mécanismes permettant de manipuler les éléments OWL,
- la mise en place d'inférences à partir d'une représentation qui implante la sémantique formelle du langage,
- la sérialisation correspondant à la production de syntaxe OWL à partir de structures de données ou représentations internes.

L'API est composée de quatre principaux paquetages : le paquetage *modèle* modélisant les différentes structures de données utilisées pour représenter les éléments OWL, le paquetage *io* définissant des fonctions de manipulation sur les éléments, le paquetage *change* gérant les modifications faites sur l'ontologie, le paquetage *inference* décrivant les mécanismes applicables à l'ontologie. L'API est conçue sur la base d'interfaces java permettant des implantations alternatives en fonction des besoins requis par les applications. Une implantation simple de l'API est disponible[6]. Cette implantation fournit les mécanismes de base dont nous avons besoin : structure de représentation des concepts, instances, relations hiérarchiques et associatives, caractéristiques des relations, manipulations, modifications et ajouts de ces éléments dans l'ontologie. Elle permet de plus le chargement d'ontologies volumineuses en mémoire virtuelle en quelques minutes. Nous l'avons donc utilisée pour implanter la première couche de notre prototype. L'optimisation de l'implantation pourrait cependant être envisagée, par exemple par le stockage de façon persistante de l'ontologie dans une base de données. Notons que parallèlement au développement de cette API, l'API Jena[7] a été également proposée pour permettre la manipulation d'ontologies OWL. Bien qu'aujourd'hui cette API offre des fonctionnalités plus élaborées que l'API OWL (stockage persistant de l'ontologie dans une base MySQL, PostgreSQL ou Oracle par exemple), au moment de la conception de notre prototype elle n'était pas disponible.

3.2 Interface

L'interface que nous avons développée vise à rendre les ontologies manipulées par le prototype visualisables et navigables par un utilisateur.

3.2.1 Visualisation d'ontologie

La visualisation d'une ontologie par un utilisateur implique la représentation graphique des éléments de l'ontologie l'intéressant. Ces éléments sont les concepts, les relations entre concepts, les instances, les relations entre instances ainsi que les labels associés à ces différents éléments. Afin de les représenter graphiquement, la bibliothèque Jgraph[8] est utilisée. Une

[6] http://owl.man.ac.uk/api.shtml

[7] http://www.hpl.hp.com/semweb/jena.htm

[8] http://www.jgraph.com/

méthode d'affichage d'ontologie à partir de cette librairie a été développée dans le cadre d'un stage réalisé par un étudiant d'IUP [Cardonner, 2004].

Pour différencier les éléments de l'ontologie, un code de représentation (couleur et forme) leur a été associé. La figure 8.2 présente une capture d'écran de l'interface permettant de visualiser une ontologie à partir des concepts qu'elle contient. Les concepts sont représentés par des rectangles pleins jaunes. Le label principal est inscrit dans ce rectangle pour désigner le concept, les autres labels sont affichés dans des boîtes au-dessous du label principal. Les instances de concepts sont représentées par des rectangles blancs au contour jaune. Dans le cadre de notre étude seules les ontologies de tâche possèdent des instances (cf figure 8.3). Les relations associatives entre concepts et entre instances sont symbolisées par des flèches au-dessus desquelles figure le nom de la relation. Dans un souci de clarté, la relation « est un » est la seule relation non labellisée. Afin de ne pas surcharger les ontologies par l'ensemble des relations définies entre concepts et instances dans l'ontologie, l'interface permet de sélectionner les relations associatives à afficher. Bien que cette fonctionnalité permette de clarifier la visualisation d'une ontologie, elle n'est intégrée à notre connaissance dans aucun outil de visualisation d'ontologie.

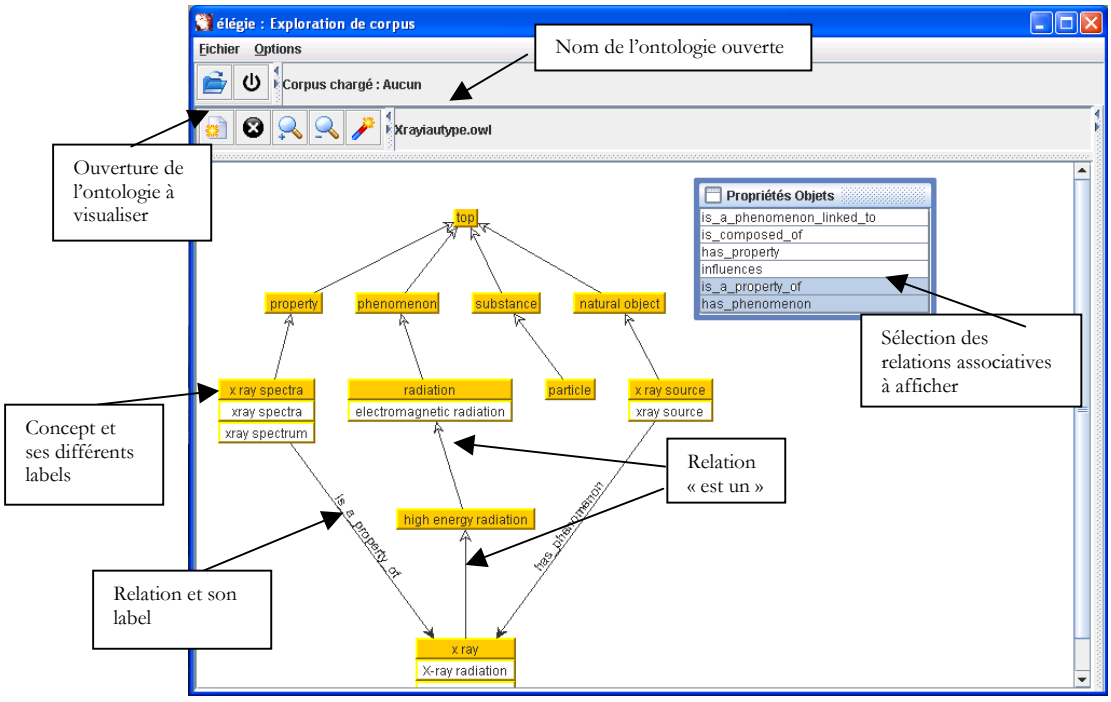


Figure 8.2 Interface de visualisation d'ontologies au niveau des concepts

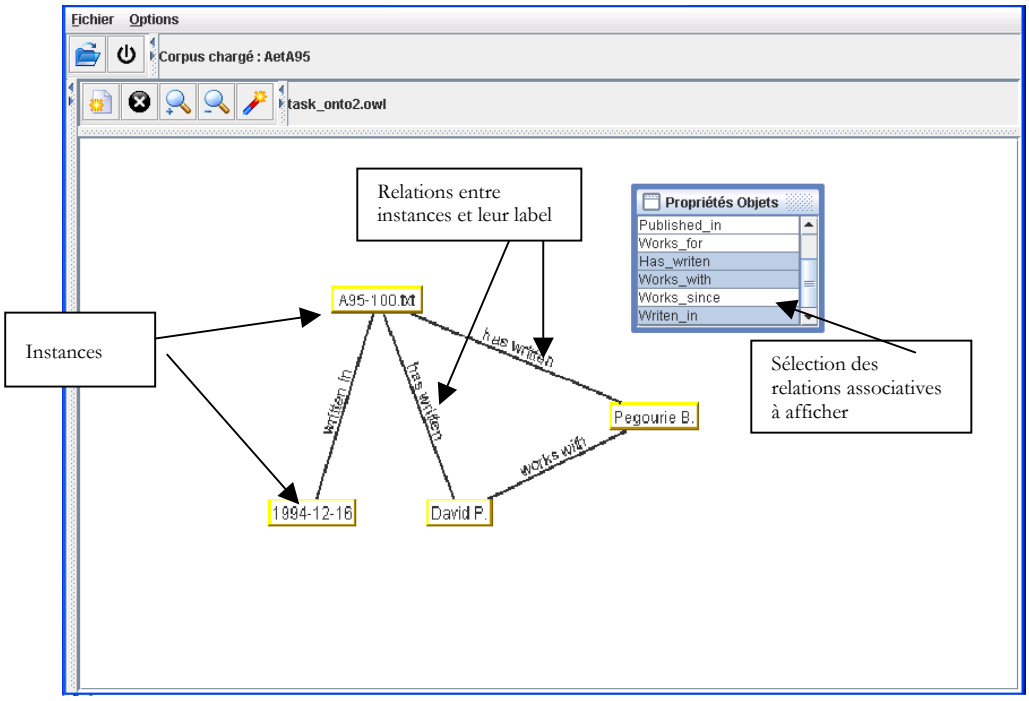


Figure 8.3 Interface de visualisation d'ontologies au niveau des instances

3.2.2 Navigation dans une ontologie

La navigation dans une ontologie est indispensable dans le cas des ontologies volumineuses qui ne peuvent pas être visualisées en intégralité sur un écran. Elle permet en effet à l'utilisateur de parcourir l'ontologie en fonction des éléments qu'il recherche.

Dans le cas de notre interface, la navigation se fait à partir de plusieurs fonctionnalités.

3.2.2.1 Choix du concept du départ de la navigation

La première fonctionnalité est le choix du concept de départ de la navigation.

Ce concept peut tout d'abord être choisi à partir de la spécification d'un label. L'ensemble des concepts définis à partir de ce label est alors proposé à l'utilisateur qui en choisit un pour commencer sa navigation.

L'utilisateur peut aussi accéder à un concept à partir de listes présentant les concepts accessibles aux différents niveaux hiérarchiques de l'ontologie. Par exemple, l'utilisateur peut parcourir l'ontologie à partir de sa racine (top). Il peut aussi parcourir la liste des concepts au niveau 1 et ainsi accéder à l'ensemble des types abstraits disponibles. Ces cas de figures sont présentés dans la figure 8.4. Les autres niveaux sont également accessibles.

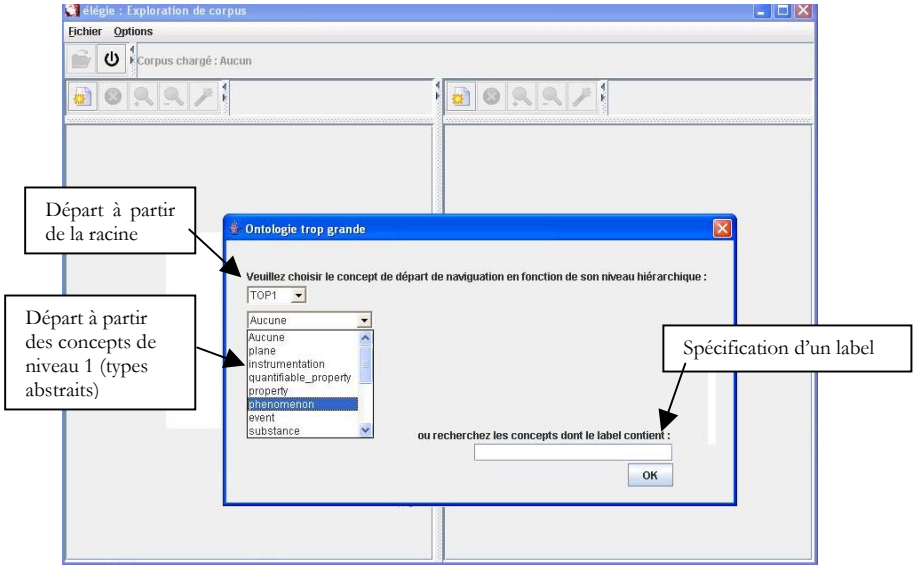


Figure 8.4 : Choix du concept départ de la navigation

3.2.2.2 Visualisation du contexte d'un concept

La deuxième fonctionnalité associée à la navigation est la visualisation du contexte d'un concept. Le contexte est représenté par l'ensemble des relations définies sur le concept (concepts père et fils, concepts liés par des relations associatives, …). Cette fonctionnalité est représentée dans la figure 8.5. L'utilisateur peut ensuite naviguer d'un concept à un autre à partir de leur contexte commun (en les sélectionnant).

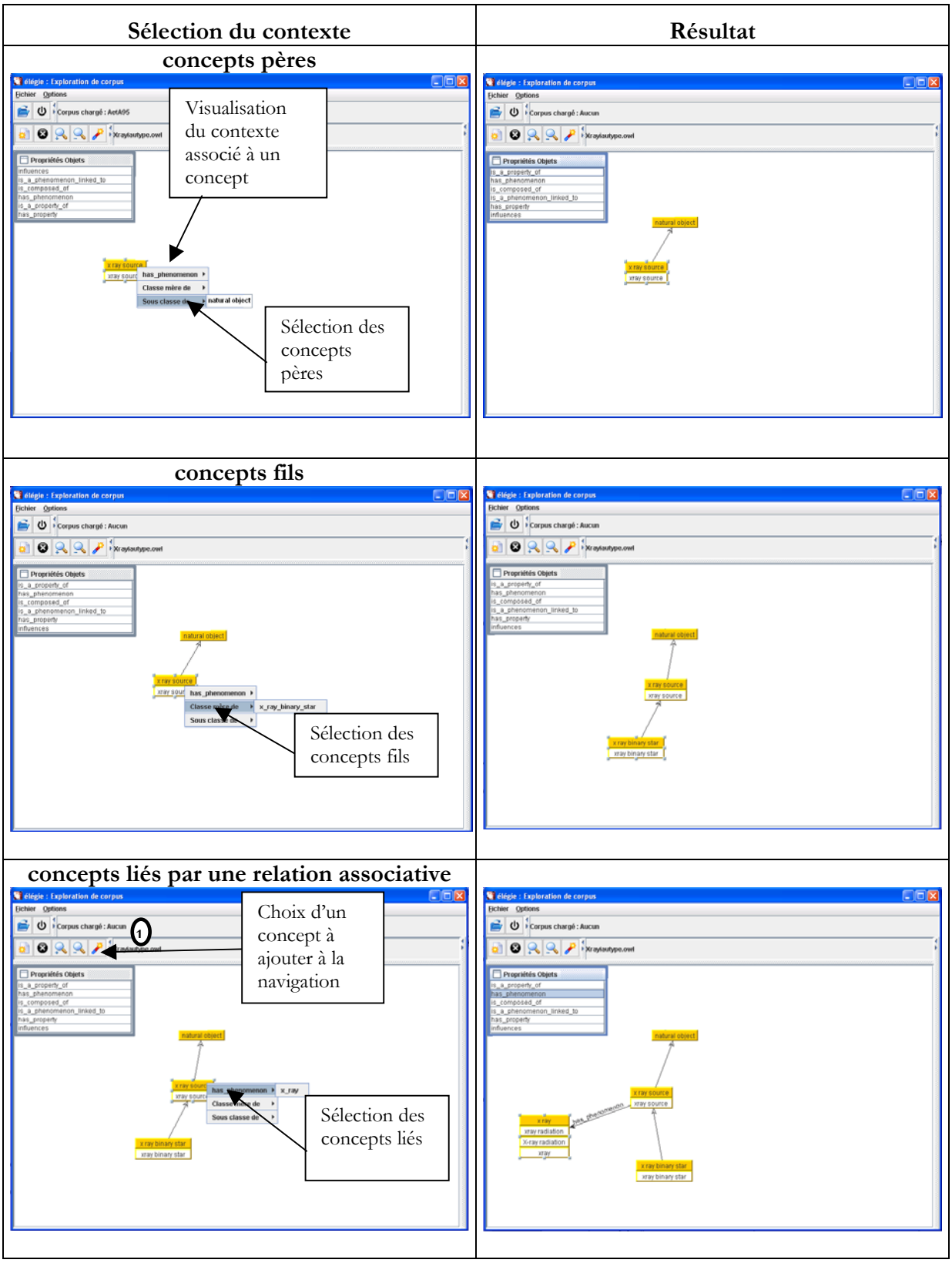


Figure 8.5 Navigation dans le contexte d'un concept

3.2.2.3 Ajouter un nouveau concept à la navigation

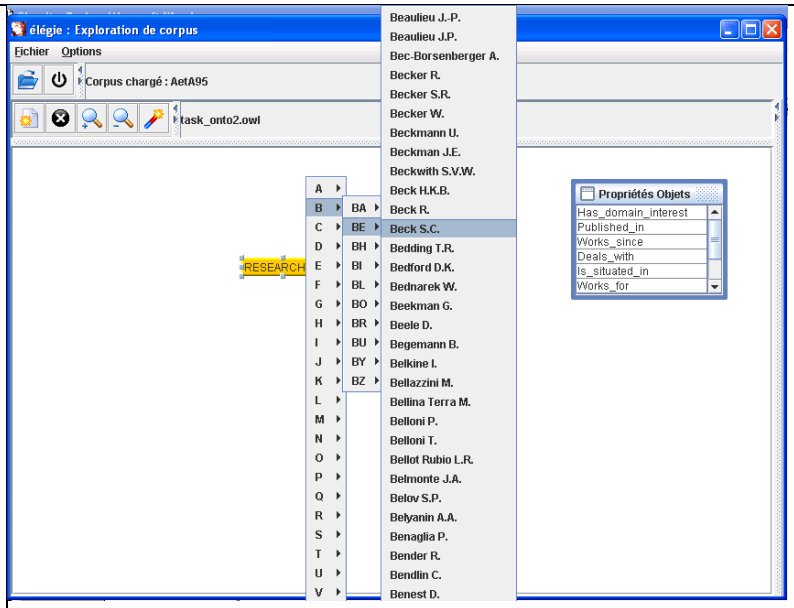
Une troisième fonctionnalité permet d'afficher d'autres concepts de l'ontologie. Les techniques utilisées sont alors celles présentées dans la section 3.2.2.1 (spécification du label, sélection par niveau de hiérarchie). L'icône (1) (cf. figure 8.5) permet l'accès à cette fonctionnalité.

3.2.2.4 Visualisation des instances

La visualisation des instances est un autre aspect du contexte associé à un concept (cf figure 8.6a). Comme pour les concepts, il est possible de naviguer dans le contexte associé à l'instance (cf figure 8.6b).

a-

Visualisation des instances correspondant au concept « *Researcher* »



b-

Visualisation du contexte de l'instance « *Beck S.C* »

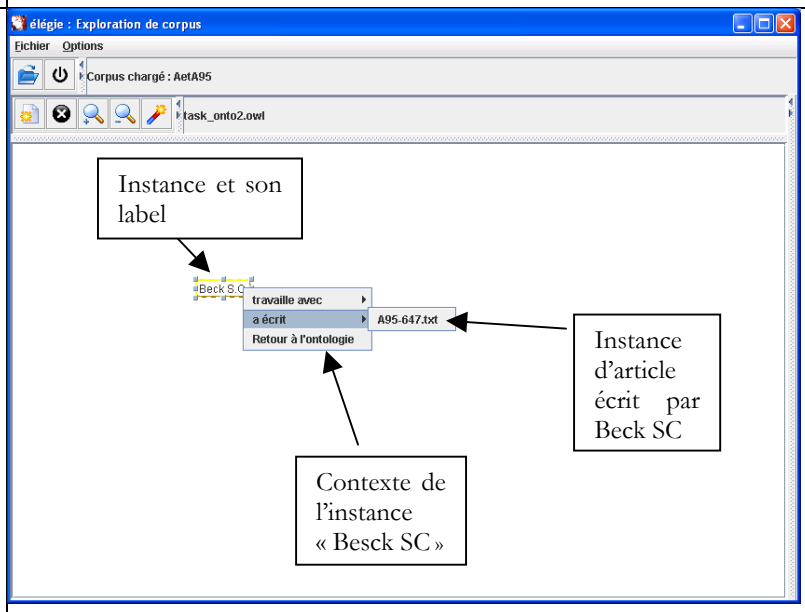


Figure 8.6 Navigation dans le contexte d'une instance

Ces mécanismes sont utiles lorsque l'ontologie est de grande taille ; dans le cas d'ontologies de petite taille (moins de 30 concepts), l'ontologie complète est affichée directement.

3.3 Classes java implantant l'interface de navigation

Si les classes java de manipulation de bas niveau des ontologies a été reprise de l'API OWL, nous avons implanté l'ensemble des classes nécessaires à l'interface de navigation.

L'ensemble de l'interface de navigation a nécessité le développement de 32 classes. Parmi celles-ci, nous décrivons plus spécifiquement les classes permettant l'affichage des concepts et des instances d'une ontologie.

Paquetage utilitaires.graphes
Class AffichageGraphe
Attributs
protected org.jgraph.JGraph **myGraph**
Jgraph correspondant à l'ontologie
protected org.jgraph.graph.GraphModel **myModel**
Modèle du Jgraph
protected org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à afficher
…
Constructeur
AffichageGraphe(java.lang.String cote)
Crée un graphe des concepts appartenant à l'ontologie. En paramètre : le côté de l'interface où se trouve l'ontologie
Méthodes
void **afficher**()
Affiche le graphe correspondant à l'ontologie
void **arrangerGraphe**()
Arrange l'ordonnancement des cellules dans le graphe
java.awt.Dimension **calculerPlacement** (org.jgraph.graph.DefaultGraphCell groupe, java.awt.Dimension d)
Calcule le placement d'une cellule du graphe. Retourne les dimensions de la cellule
…

Paquetage utilitaires.graphes
Class AffichageGrapheInstances
Attributs
protected org.jgraph.JGraph **myGraph**
Jgraph correspondant à l'ontologie
protected org.jgraph.graph.GraphModel **myModel**
Modèle du Jgraph
protected org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à afficher
…
Constructeur
AffichageGrapheInstances(java.lang.String cote, java.lang.String inst, java.lang.String type)
Crée un graphe à afficher à partir des instances d'une ontologie. En paramètre : le côté de l'interface où se trouve l'ontologie, ainsi que le nom de l'instance au départ de la navigation et le concept auquel elle se rapporte
Méthodes
void **ajouterInstance**(java.lang.String inst, java.lang.String type)
Ajoute une instance au graphe
void **ajouterLiens**(java.lang.String inst, java.lang.String type)
Ajoute au graphe les liens connus pour l'instance de nom inst se rapportant au concept type
….

Figure 8.7 Extrait de la spécification des classes affichage_graphe et affichage_graphe_instance.

4 Analyse de l'adéquation d'ontologies à un corpus

La couche d'analyse de l'adéquation d'une ontologie de thème à un corpus vise à valider le choix d'une ontologie de thème pour la représentation du contexte associé à un corpus. Elle comporte l'implantation des méthodes présentées dans le chapitre 6 et une interface permettant d'en visualiser les résultats.

4.1 Implantation

L'implantation de la couche d'analyse de l'adéquation est réalisée par la programmation en java des méthodes présentées dans le chapitre 6. Elle est programmée par l'intermédiaire de classes organisées suivant deux paquetages.

Le premier paquetage regroupe les traitements relatifs à l'analyse lexicale. Une de ces classes permet d'accéder au lexique du corpus (charger_corpus) en traitant les fichiers relatifs à l'analyse syntaxique du corpus réalisé par Syntex. Cette classe effectue de plus la correspondance entre le lexique du corpus et le lexique de l'ontologie. Elle est présentée dans la figure 8.8.

utilitaires.adequation_lexicale
Class **ChargerCorpus**
Attributs
(package private) java.lang.String **corpus**
Emplacement des fichiers en sortie de Syntex
(package private) java.io.BufferedReader **lecteurAvecBuffer**
Tampon utilisé pour la lecture des fichiers Syntex
(package private) java.lang.String **cote**
Côté de l'interface où se trouve l'ontologie
(package private) org.semanticweb.owl.model.OWLOntology **onto**
Ontologie traitée
(package private) java.util.HashMap **labconcepts**
hachage des termes extraits par Syntex et des concepts auxquels ils correspondent dans l'ontologie
Constructeur
ChargerCorpus(java.lang.String corpus, java.lang.String cote)
Recherche des termes extraits par Syntex qui sont dans le lexique de l'ontologie et détermination des concepts correspondant.
En paramètre : emplacement des fichiers résultant de Syntex et côté de l'interface dans lequel a été ouverte l'ontologie

Méthodes
java.util.HashMap **label_concept**(java.util.HashMap labels_onto, java.lang.String terme)
Renvoie pour un terme donné l'ensemble des concepts de l'ontologie auxquels il correspond
org.semanticweb.owl.model.OWLClass **desamb_label**(java.util.HashMap labels_terme)
Renvoie pour un terme label de plusieurs concepts, le concept effectivement référencé dans le document
…

Figure 8.8 Extraits de la spécification d'une classe réalisant l'analyse lexicale

Le deuxième paquetage implante l'analyse conceptuelle. Il est composé de plusieurs classes permettant le calcul de proximité entre concepts (ProxJing), de la représentativité d'un concept et de la représentativité d'une ontologie (PowerJing). Un extrait de la spécification de ces classes est présenté dans la figure 8.9.

utilitaires. adequation_conceptuelle
Class ProxJing
Attributs
org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à traiter
java.lang.String **corpus**
Corpus considéré
java.util.HashMap **classefreqdep**
Fréquence des concepts dans le corpus
…
Constructeur
ProxJing(java.lang.String cote, java.lang.String corpus)
Calcule la similarité entre tous les concepts de l'ontologie chargée dans le côté de l'interface à partir du corpus

Méthodes
static chemin **pluscourtchemin** (org.semanticweb.owl.model.OWLClass conc1, org.semanticweb.owl.model.OWLClass conc2,)
Calcule le plus court chemin entre deux concepts
static double **poids_chemin**(chemin ch, org.semanticweb.owl.model.OWLClass c1, org.semanticweb.owl.model.OWLClass c2, java.util.HashMap classproba)
Calcule le poids du plus court chemin entre deux concepts à partir de leur contenu en information dans le corpus
…

utilitaires. adequation_conceptuelle
Class PowerJing
Attributs
org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à traiter
java.lang.String **corpus**
Corpus considéré
java.util.double[][]Mat=new double[NBconcept][NBconcept]
Stocke la similarité entre tous les concepts de l'ontologie
…
Constructeur
ProwerJing(java.lang.String cote, java.lang.String corpus)
Calcule le pouvoir représentatif de tous les concepts de l'ontologie chargée dans le côté de l'interface à partir du corpus

Méthodes
static double **getProx**((org.semanticweb.owl.model.OWLClass conc1, org.semanticweb.owl.model.OWLClass conc2,)
Récupère la proximité sémantique entre deux concepts
static double **power** (org.semanticweb.owl.model.OWLClass c,)
Calcule le pouvoir représentatif d'un concept

static double **power_onto();**
Calcule le pouvoir représentatif de l'ontologie traitée
…

Figure 8.9 Extraits de la spécification des classes réalisant l'analyse conceptuelle

4.2 Interface

L'interface associée à la couche d'analyse de l'adéquation permet de sélectionner une ou deux ontologies de thème et un corpus. Les ontologies sont affichées à partir des techniques décrites dans la section 3. Elle présente la visualisation des résultats de l'analyse. Par son intermédiaire, l'utilisateur peut interpréter les résultats de l'adéquation entre une ontologie et un corpus. Il peut ainsi valider le choix d'une ontologie ou choisir entre deux ontologies existantes par la comparaison des résultats obtenus sur les deux ontologies pour un même corpus. Afin de permettre la comparaison entre deux ontologies, l'interface est divisée en deux fenêtres. A partir

d'une boite à outils associée à cette interface, l'utilisateur peut ouvrir deux ontologies différentes et charger un corpus. L'interface présente à la fois les résultats de l'analyse conceptuelle et ceux de l'analyse lexicale. Dans chacune des fenêtres, les ontologies sont visualisables et navigables. Des fenêtres d'information leur sont associées afin de présenter les détails des deux analyses.

4.2.1 Visualisation des résultats de l'analyse conceptuelle

Une fois les ontologies et le corpus choisis, l'interface présente à l'utilisateur le résultat de l'analyse conceptuelle. Cette analyse vise à évaluer la représentativité de chaque concept ainsi que la représentativité de l'ontologie pour le corpus considéré. La figure 8.10 présente une copie d'écran relative aux résultats de la comparaison de deux ontologies sur un même corpus.

Afin que la représentativité d'un concept soit visuellement interprétable par l'utilisateur, les concepts sont présentés à partir d'un code de couleurs associées à chaque tranche de valeur de la représentativité. Des couleurs sont définies par défaut par le système mais elles peuvent être modifiées par l'utilisateur.

La partie (1) de figure 8.10 présente un exemple de code de couleur. Lorsque l'utilisateur sélectionne un concept, il a accès à la valeur exacte de la représentativité d'un concept. Cette valeur s'affiche dans la fenêtre d'information (partie (2) de la figure). Le pouvoir représentatif global de l'ontologie est également présenté à l'utilisateur, il correspond à la partie (3) de la figure. Le nombre de concepts de l'ontologie non retrouvés dans le corpus est aussi intégré à la fenêtre d'information (partie (4) de la figure).

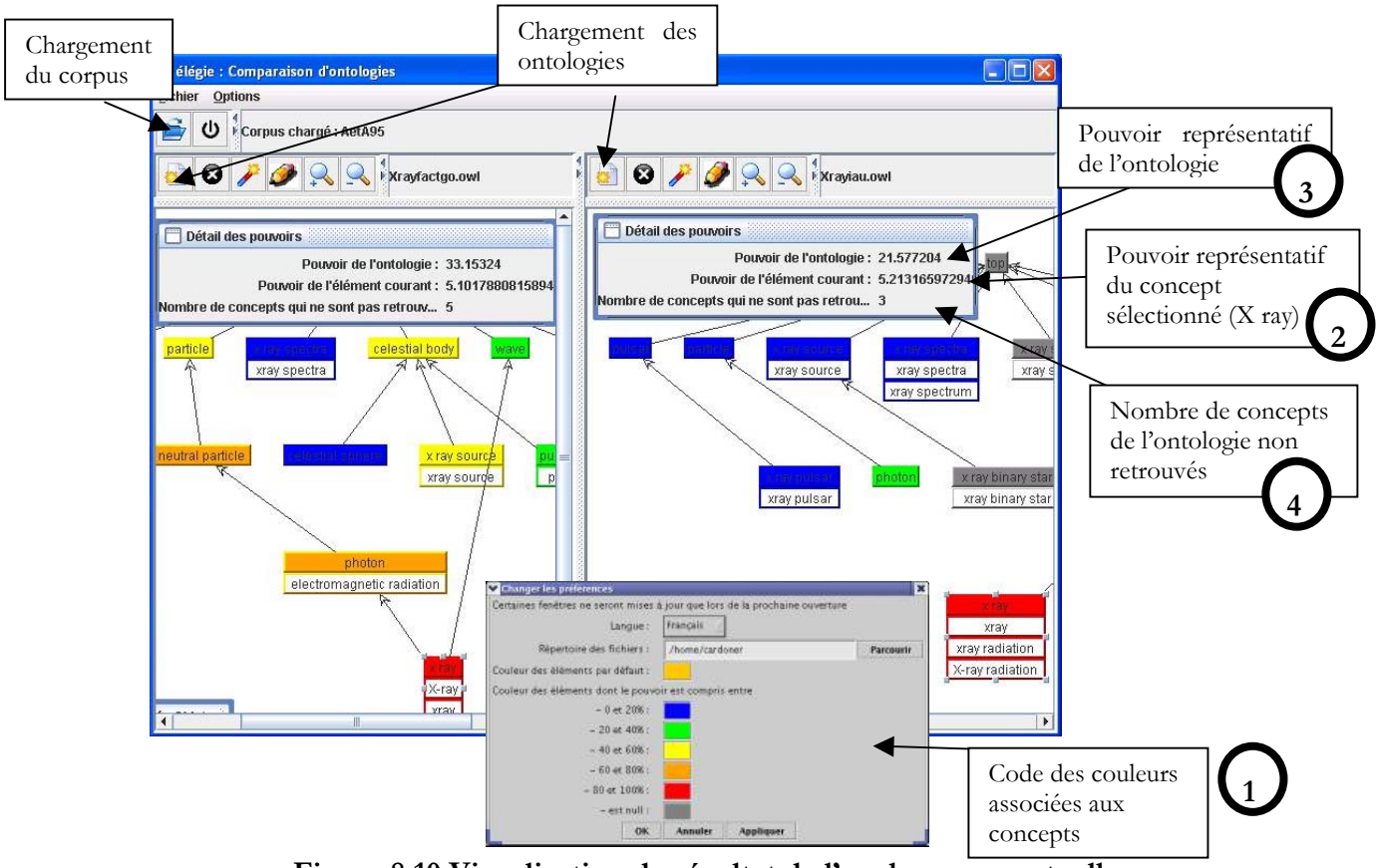


Figure 8.10 Visualisation du résultat de l'analyse conceptuelle

4.2.2 Visualisation des résultats de l'analyse lexicale

L'interface présente également les résultats de l'analyse lexicale. Une partie de ces résultats sont visualisables par l'utilisateur par le clic sur un icône dédié de la barre d'outils.

Une fenêtre d'information présentant les termes spécifiques et les génériques du corpus non retrouvés dans l'ontologie apparaît alors dans l'interface. Un exemple de ce type de résultat est présenté dans la figure 8.11.

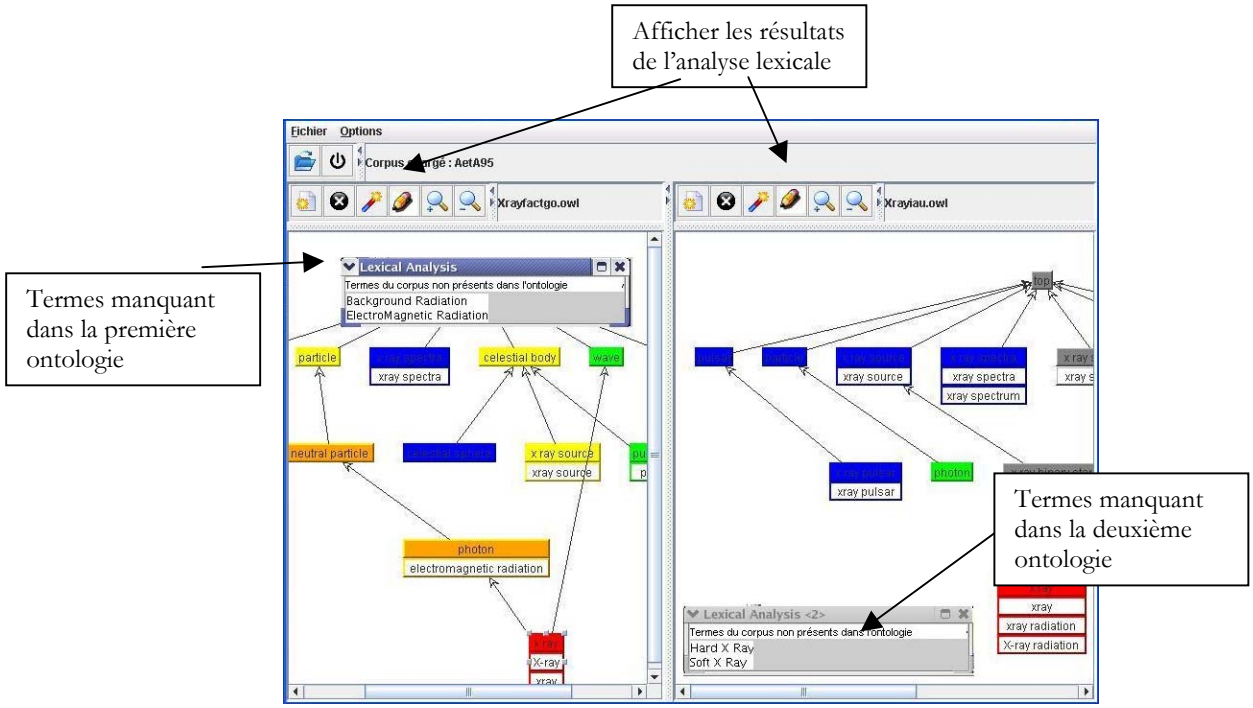


Figure 8.11 Visualisation des termes du corpus manquant dans l'ontologie

Le contexte d'un concept est complété par la liste des documents qui le référencent. L'utilisateur a alors accès aux documents référençant les concepts. Les labels du concept retrouvés lui sont indiqués par leur mise en évidence dans le contenu du document. La figure 8.12 présente cet aspect. Grâce à ce mécanisme, l'utilisateur peut vérifier la pertinence des concepts retrouvés dans les documents et évaluer le mécanisme de désambiguïsation mis en place.

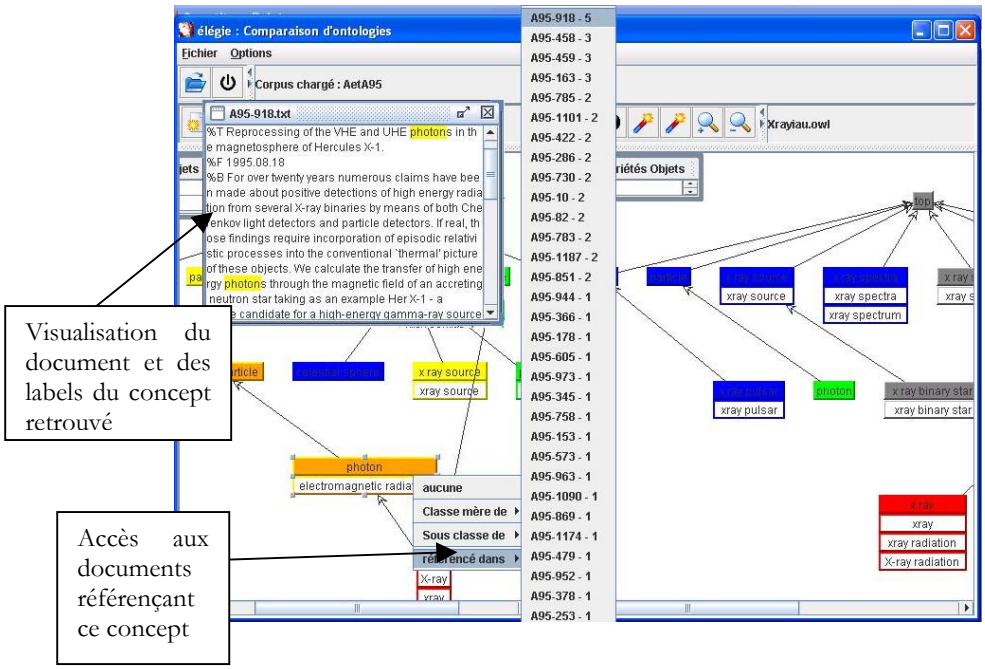


Figure 8.12 Visualisation des documents du corpus traitant d'un concept

5 Intégration du contexte dans le traitement du corpus

La dernière couche du système intègre le contexte représenté par l'ontologie du thème et l'ontologie de la tâche au traitement du corpus. Cette couche comprend l'implantation de l'annotation des documents suivant les deux ontologies et la définition d'une interface permettant à l'utilisateur d'accéder aux documents suivant le contexte de la recherche.

5.1 Implantation

L'implantation de cette couche comprend la mise en place de l'indexation sémantique à partir de l'ontologie de thème et l'extraction des instances à partir de l'ontologie de tâche.

5.1.1 Indexation sémantique à partir de l'ontologie de thème

L'implantation de l'indexation sémantique de granules documentaires par l'ontologie du thème est réalisée par la programmation en java des mesures présentées dans la section 3.1.1 du chapitre 4. Pour un granule donné, différentes classes permettent d'extraire les concepts référencés ainsi que de les classer à partir de leur représentativité sémantique et statistique.

La classe indexation_sémantique permet d'effectuer cette indexation (cf figure 8.13).

utilitaires. index
Class Indexation_semantique
Attributs
org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à traiter
java.lang.String **corpus**
Corpus considéré
…
Constructeur
Indexation_semantique(java.lang.String cote, java.lang.String corpus)
Indexe le corpus de documents à partir de l'ontologie chargée dans le côté cote de l'interface

Méthodes
java.util.HashSet GetConceptDocument(java.lang.String Contenu_document)
Recupère les concepts retrouvés dans un document
Java.util.double GetRepresentativité_sem(java.lang.String Contenu_granule, org.semanticweb.owl.model.OWLClass c)
Calcule la représentativité sémantique du concept c dans le granule contenu_granule
Java.util.double GetRepresentativité_stat(java.lang.String Contenu_granule, org.semanticweb.owl.model.OWLClass c)
Calcule la représentativité statistique du concept c dans le granule contenu_granule
…

Figure 8.13 Extrait de la spécification de la classe d'indexation_sémantique

5.1.2 Extraction des instances de l'ontologie de tâche

L'extraction des instances de l'ontologie de tâche est réalisée par les techniques décrites dans la section 3.1.2 du chapitre 4. Leur implantation est divisée en deux classes distinctes.

La première vise à extraire les méta-données explicitées dans les documents et à créer les instances correspondant aux concepts de l'ontologie de tâche.

La seconde identifie et stocke les instances de l'ontologie de tâche présentes dans l'ontologie de thème à partir du résultat de l'indexation sémantique sur les granules considérés.

La classe Charger_Instances permet de mettre en place ce mécanisme (cf. figure 8.14).

L'ensemble des instances est représenté dans le modèle du domaine associé à la tâche. Ce modèle est stocké dans la base de données présentée dans la section suivante.

utilitaires. index
Class Charger_Instances
Attributs
org.semanticweb.owl.model.OWLOntology **onto**
Ontologie à traiter
java.lang.String **corpus**
Corpus considéré
…
Constructeur
Charger_Instances (java.lang.String cote, java.lang.String corpus)
Extrait les instances de l'ontologie chargée dans le côté de l'interface dans les documents du corpus

Méthodes
void **creerInstance** (java.lang.String nom, org.semanticweb.owl.model.OWLClass concept_corresp, org.semanticweb.owl.model.change.ChangeVisitor visitor, org.semanticweb.owl.model.OWLDataFactory fact, org.semanticweb.owl.model.change.OntologyChange oc)
Une instance est ajoutée à l'ontologie

Figure 8.14 Extrait de la spécification de la classe charger_instance

5.1.3 Stockage des annotations

Afin de gérer et d'optimiser l'accès aux annotations des documents par les deux ontologies, les annotations doivent être stockées par un mécanisme adapté. Le stockage doit en effet, permettre de gérer les annotations relatives à un ensemble volumineux de granules mais aussi être capable d'intégrer de nouveaux documents ajoutés à la collection. Il est de plus nécessaire, par souci d'interactivité entre le système et l'utilisateur, que les annotations puissent être accessibles rapidement.

Pour permettre un accès rapide aux instances de l'ontologie de tâche correspondant aux annotations et ne pas augmenter le temps de chargement du système d'exploration, nous proposons de les stocker de façon persistante dans une base de données relationnelle. La modélisation de la base est établie pour optimiser l'accès aux informations élaborées à partir des deux ontologies. Le diagramme de classes UML de la base de données utilisée est représenté dans la figure 8.15.

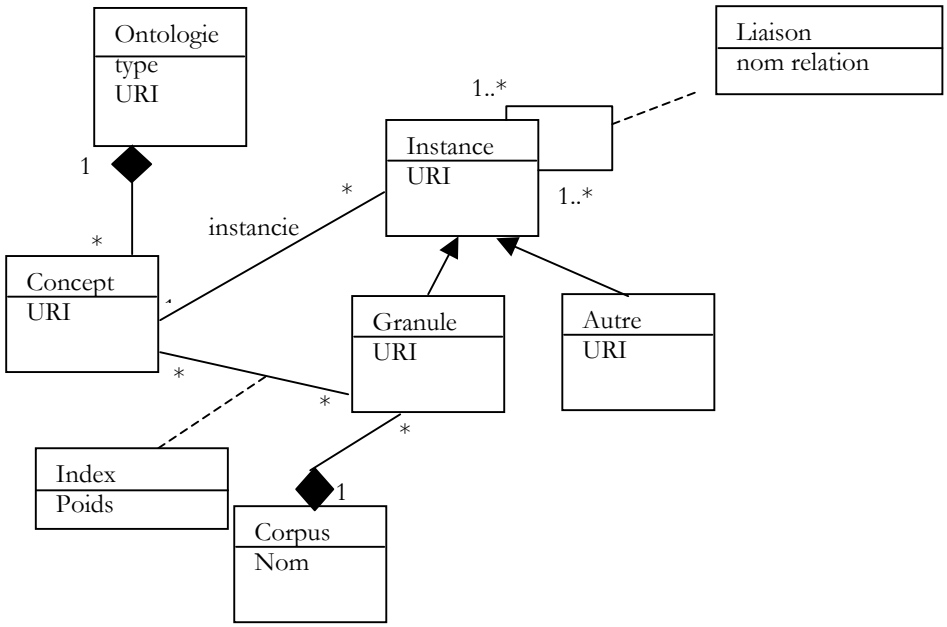


Figure 8.15 Diagramme de classes UML de la base de données relationnelle utilisée pour stocker les annotations

Les objets *instance* correspondent à l'ensemble des instances de l'ontologie de tâche. Ces objets sont regroupés en deux sous-classes : les granules pour lesquels une indexation sémantique est réalisée (ce qui justifie la présence de la classe d'association *Index*) et les autres instances explicitement présentes dans les documents. Un concept appartient à une ontologie. Cette ontologie est soit de thème soit de tâche en fonction de son *type*. Le concept qui fait le lien entre les deux ontologies est un objet de la classe *concept* pour l'ontologie de thème et une occurrence de la classe d'association *Index* pour l'ontologie de tâche.

La base de données est implantée sous Mysql.

5.2 Interface de visualisation des données

Une interface d'accès aux annotations issues des deux ontologies a été développée. Une copie d'écran de cette interface est présentée dans la figure 8.16. L'interface intègre les boîtes à outils permettant d'ouvrir les deux ontologies et de charger un corpus qui ont été présentées précédemment.

5.2.1 Visualisation des ontologies

Cette interface permet de visualiser à la fois l'ontologie de tâche (ici la veille) et l'ontologie de thème (ici l'astronomie). L'exploration du corpus peut donc s'effectuer à partir de *deux types de connaissances* : la connaissance liée à la tâche d'exploration et la connaissance liée au domaine du thème. Pour cela, l'écran est divisé en deux fenêtres, la fenêtre de droite étant dédiée à la visualisation de l'ontologie du domaine de la tâche, la fenêtre de gauche à celle de l'ontologie du domaine du thème.

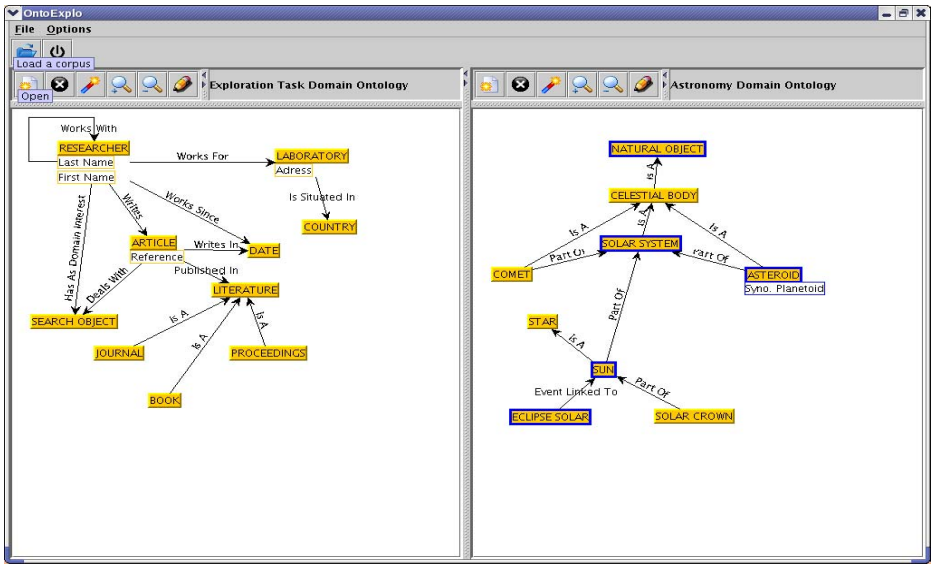
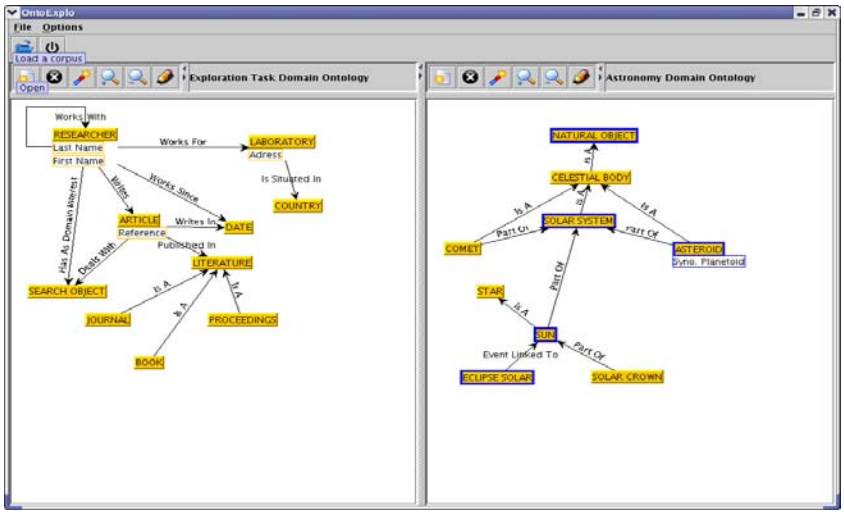


Figure 8.16 Interface d'exploration du corpus

5.2.2 Niveaux d'exploration

L'interface permet *deux niveaux d'exploration* du corpus. Le premier niveau correspond aux concepts des ontologies. Ainsi l'utilisateur peut avoir une vue d'ensemble sur la collection et la connaissance qui lui est associée. Le second niveau s'effectue à partir des instances des concepts de l'ontologie de tâche. Le passage du premier niveau au deuxième est réalisé en cliquant sur un concept. Ceci permet de rendre accessible à l'utilisateur les détails sur le contenu de la collection (cf figure 8.16 a et b)

a-Au niveau concept



b-Au niveau instance

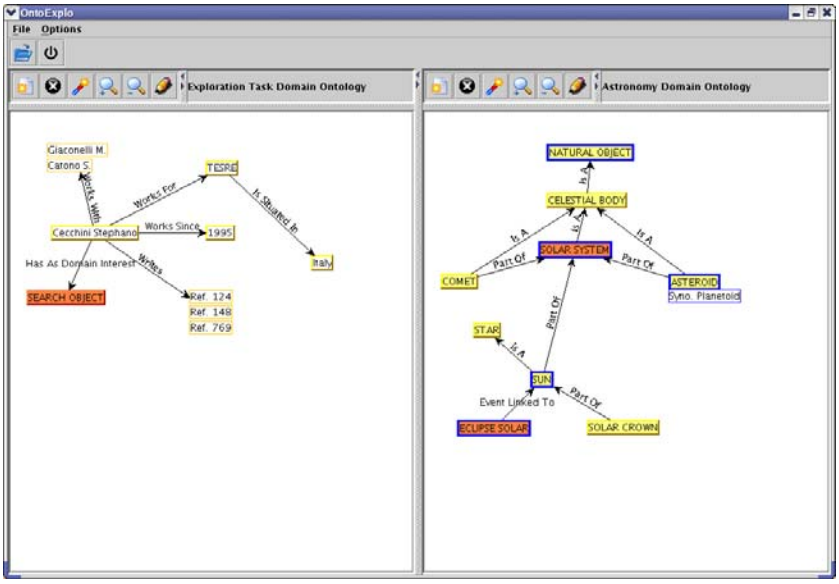


Figure 8.16 Niveaux d'exploration

Les deux sections suivantes décrivent les deux types d'explorations accessibles à partir de l'interface : l'exploration à partir de l'ontologie du domaine de la tâche et l'exploration à partir de l'ontologie du domaine du thème, en précisant l'intérêt des deux niveaux d'exploration (concept-instance).

5.3 Exploration à partir de l'ontologie du domaine de la tâche

Lorsque l'utilisateur est intéressé par les méta-données relatives à la tâche qu'il entreprend, il choisit d'explorer le corpus à partir de l'ontologie de tâche. L'utilisateur découvre alors tous les acteurs et les méta-données du corpus relatifs à la tâche (partie gauche de la figure 8.9). Le concept *Search Object* de l'ontologie de la tâche de veille est un lien conceptuel entre les deux ontologies. Il fait référence à un concept de l'ontologie du thème de l'astronomie.

En naviguant dans l'ontologie de la tâche, l'utilisateur peut parcourir les différentes instances créées à partir de la collection pour ce concept en cliquant sur le label principal du concept. Dans la figure 8.17, le concept *Researcher* a été sélectionné et l'ensemble des instances (nom des chercheurs) est affiché. L'utilisateur peut focaliser son attention sur une instance particulière en la sélectionnant. Dans la figure 8.18, l'utilisateur a sélectionné *Cecchini S.* En sélectionnant l'instance qui est l'objet de sa requête dans l'ontologie du domaine de la tâche (e.g. *Researcher* Cecchini S.), l'utilisateur visualise les relations connues entre cette instance et les autres instances de l'ontologie : chercheurs avec qui il travaille : Giaconelli M.,…, publications : ref 124, 148, 789, laboratoire d'affiliation : TESRE…

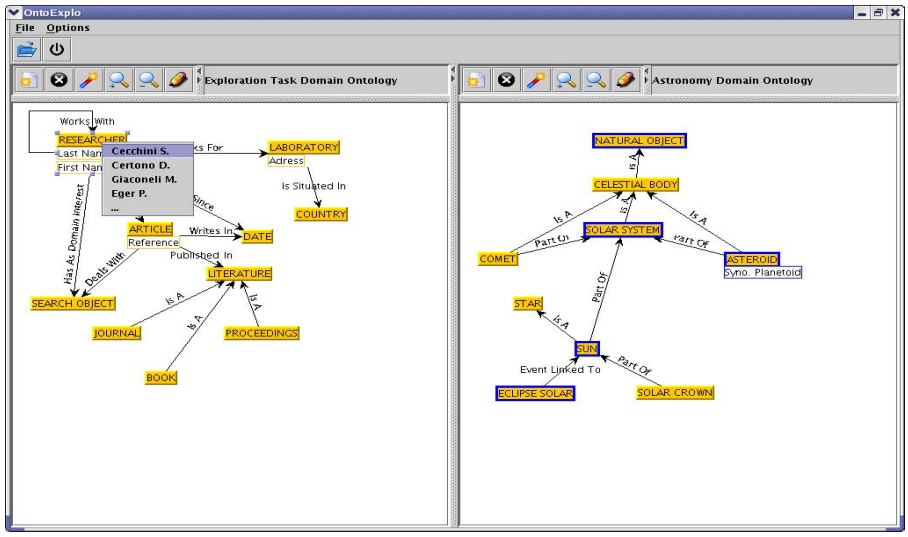


Figure 8.17. Visualisation des instances du concept Researcher de l'ontologie du domaine de la tâche

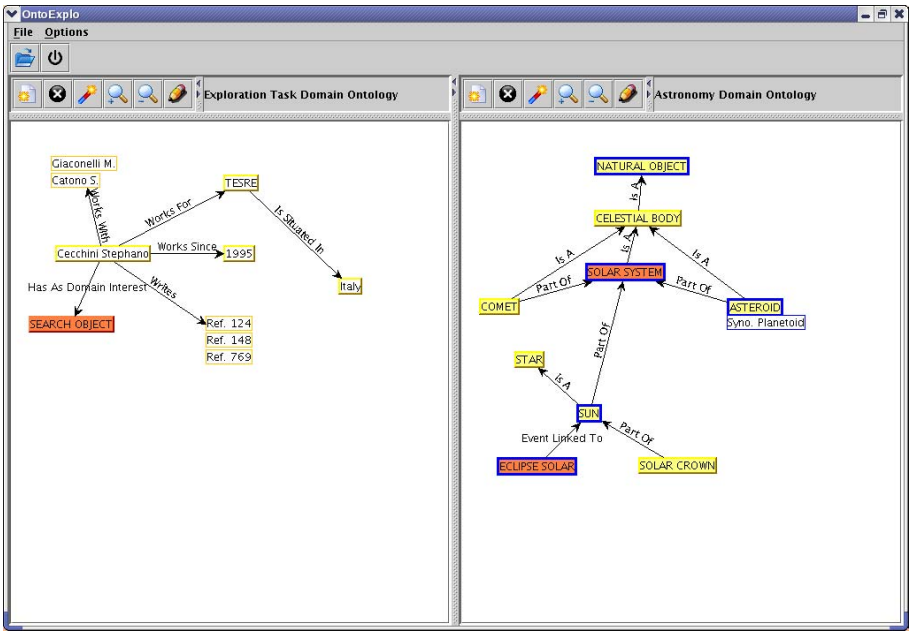


Figure 8.18. Visualisation de la connaissance établie pour une instance du concept *Researcher* de l'ontologie du domaine de la tâche

Le lien entre les deux ontologies est aussi présenté à l'utilisateur. Dès sélection du concept, les instances de concept sont colorées. Par exemple, dans la figure 8.18, le concept *Research Object* a été sélectionné. Cela a entraîné la coloration foncée de celui-ci dans l'ontologie de tâche (partie gauche de l'interface) et celle des concepts correspondants dans l'ontologie de thème (partie droite de l'interface). Dans la figure 8.18, les instances du concept *research object* liées au chercheur Cecchini S. sont *solar system* et *eclipse* ; ce chercheur a donc comme centres d'intérêt ces deux concepts du domaine de l'astronomie (cette connaissance est extraite à partir des documents de la collection). La coloration des deux concepts permet à l'utilisateur de visualiser ces concepts dans leur contexte à partir de l'échantillon de l'ontologie dans lequel ils sont présentés (autres labels pour les concepts, concepts liés à ces concepts).

Si l'utilisateur est intéressé par la visualisation d'un article, il peut sélectionner la référence de l'article, soit à partir des instances du concept *Article* (figure 8.19), soit à partir des instances des articles écrits par un auteur spécifique (figure 8.18). Comme présenté dans la figure 8.19, une fenêtre contenant l'article apparaît et l'information accessible depuis l'interface est automatiquement mise à jour. L'information connue sur l'article et utile pour la tâche apparaît dans la partie de l'écran dédiée à l'ontologie du domaine de la tâche et l'ensemble des thèmes du domaine abordés dans l'article est présenté dans la fenêtre dédiée à l'ontologie du domaine. Ces concepts sont des instances du concept *Research Object* au sein du modèle de l'article du point de vue de la tâche. L'utilisateur a ainsi toutes les informations sur ce document. Par exemple, la figure 8.19 montre que l'article de référence 124 traite de *Comet* et de *Solar system*. L'utilisateur peut ainsi rapidement évaluer si le document traite des thèmes qui l'intéressent, savoir qui a écrit l'article, où il a été publié et quand.

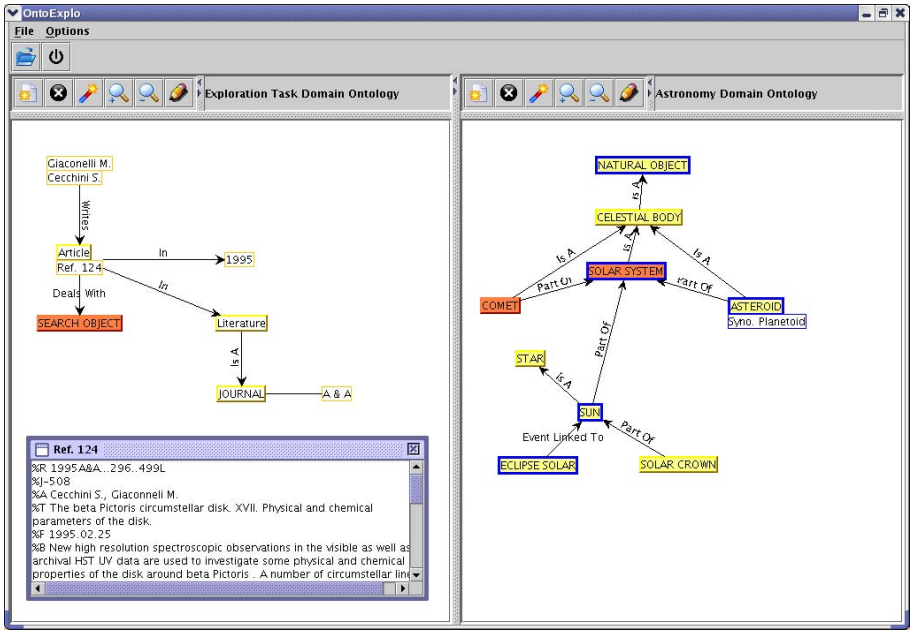


Figure 8.19. Visualisation de la connaissance établie pour une instance du concept *article* de l'ontologie du domaine de la tâche

5.4 Exploration à partir de l'ontologie du thème

L'utilisateur peut aussi choisir de naviguer au sein de l'ontologie du thème de l'astronomie. Il explore alors la collection à partir de l'ontologie du domaine abordé dans le contenu. Ce type d'exploration permet à l'utilisateur de trouver de l'information à propos d'un thème spécifique du domaine.

Cette possibilité est présentée dans la figure 8.20. Tous les concepts et les relations de l'ontologie sont affichés. Les concepts présents dans le corpus sont mis en évidence par une bordure plus épaisse. Ils peuvent être interprétés dans leur contexte puisque les relations aux autres concepts sont affichées. A partir d'un clic sur un concept, l'utilisateur peut retrouver quels sont les articles abordant ce concept ainsi que les chercheurs s'intéressant à cette thématique. Lorsqu'un chercheur est sélectionné, l'interface est automatiquement mise à jour pour afficher la connaissance élaborée pour ce chercheur (par exemple en sélectionnant *Cecchini S*, la figure 8.18 sera affichée). De la même façon, lorsque l'utilisateur est intéressé par l'accès à un article, il sélectionne sa référence (cf figure 8.20), la fenêtre est mise à jour et les informations associées pour l'article apparaissent (comme dans la figure 8.19).

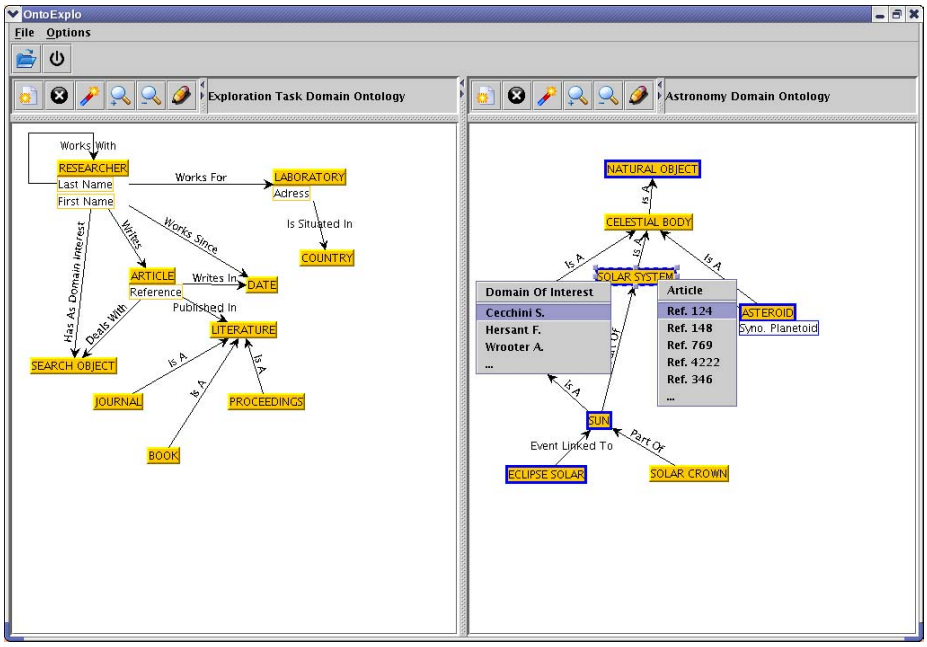


Figure 8.20 Navigation au sein de l'ontologie de l'astronomie

6 Conclusion

Le prototype OntoExplo intègre la modélisation du contexte associé à la tâche de recherche et au thème du corpus dans une activité de RI. Nous avons montré qu'il permet une navigation riche qui combine des éléments de thème et de tâche. D'autre part, la navigation à plusieurs niveaux d'abstraction propose différentes vues sur la collection. Ces vues permettent tout d'abord d'évaluer l'adéquation entre une ontologie de thème et un corpus aidant ainsi l'utilisateur dans le choix de l'ontologie de thème spécifiant au mieux le contexte associé au corpus. Parallèlement, le système offre des fonctionnalités de navigation dans des ontologies volumineuses (choix des concepts à afficher, présentation du contexte d'un concept) qui complètent cette aide à l'utilisateur. Finalement, ces mécanismes sont repris et complétés pour permettre l'accès aux informations élaborées à partir des deux ontologies et du corpus.

D'un point de vue technique, nous proposons actuellement d'optimiser l'accès aux instances et aux informations associées par l'utilisation d'une base de données (cf figure 8.15), mais de maintenir la gestion des concepts au niveau de la mémoire virtuelle à partir de la description de l'ontologie dans un fichier OWL. Une extension possible serait le stockage de l'ensemble des éléments de l'ontologie dans une base de données.

Dans notre approche, l'utilisateur reste maître du choix des ontologies. Cependant, ce choix nécessite une certaine maîtrise des outils que nous proposons ainsi qu'un temps de réalisation. Une extension de ces travaux pourrait donc être le choix automatique de l'ontologie de domaine la plus adaptée.

Conclusion générale

La RI contextuelle a pour but de replacer l'utilisateur au cœur des processus de celle-ci. L'objectif n'est plus de proposer des systèmes généralistes satisfaisant le plus grand nombre d'utilisateurs dans la plupart des cas, mais de développer des systèmes capables de répondre au besoin spécifique de l'utilisateur. Les éléments de connaissance liés à son besoin en information (ses intentions, son environnement, son expertise du domaine, ses préférences) doivent alors être explicités. La variabilité des contextes rend difficile leur modélisation et leur prise en compte dans un SRI [Harman 2004]. Même s'il n'est pas toujours explicité, le contexte est néanmoins omniprésent en RI. Certains aspects tels que le domaine de connaissances lié à un corpus sont intégrés au système depuis plusieurs années. L'utilisation de ressources de domaine n'est pas nouvelle, mais la formalisation de ces ressources a donné un nouvel élan aux recherches visant à les prendre en compte. Les thésaurus ont largement été utilisés dès le début de l'informatisation des ressources documentaires. Les systèmes de gestion documentaire intègrent d'ailleurs aujourd'hui encore ce type de ressources. Par exemple, BCDI[1], logiciel de gestion documentaire et d'interrogation utilisé dans l'Education Nationale française repose sur le thésaurus MOBIS. Les thésaurus sont également utilisés comme ressources à l'indexation manuelle, soit par des éditeurs, soit par des services en charge d'associer des mots-clés à des documents. Dans le domaine de l'astronomie, des documentalistes ont en charge d'associer des mots-clés aux publications scientifiques du domaine à partir d'un langage contrôlé ; les éditeurs en font parfois de même. L'accroissement de la mise à disposition très rapide d'informations électroniques a toutefois conduit à l'abandon de l'utilisation systématique de telles ressources. Ainsi, les premiers systèmes mis au point dans le domaine de la recherche d'information en font abstraction. L'indexation des documents repose uniquement sur le contenu des documents et l'interrogation est proposée en langage libre. SMART [Salton 1971] et OKAPI [Robertson 1976] font partie de ce type de systèmes. Pourtant, l'intérêt d'utiliser une terminologie pour classer les ressources documentaires est resté très présent. Le cas des portails et annuaires tels que Yahoo[2], en est une illustration. Cependant, le manque de formalisation de ces ressources les rend difficilement exploitables par des traitements automatiques qui sont devenus indispensables pour permettre la gestion et l'accès aux masses de données actuellement disponibles. Même si certaines recherches ont essayé de réintroduire les thésaurus ou les ressources terminologiques, en particulier pour proposer une reformulation automatique de requêtes [Voorhes 1994], ce n'est que très récemment que cette utilisation a été remise au goût du jour. Ce regain d'intérêt n'est apparu que grâce à l'émergence du support formalisé que sont les ontologies. Ces représentations spécifient la connaissance liée à un phénomène du monde. Elles définissent le sens des objets tout d'abord à travers des symboles (mots ou expressions) qui les désignent et les caractérisent, puis à travers une représentation structurée ou formelle de leur rôle dans le domaine. Ces deux niveaux de spécification permettent de les utiliser aussi bien pour la communication entre personnes qu'entre personnes et machines. Ils sont au cœur du Web sémantique [Berners-Lee 2001]. Leur intégration dans les SRI est donc possible. Par la variabilité des connaissances qu'elles peuvent représenter, elles permettent même d'envisager leur utilisation pour la modélisation d'autres aspects du contexte d'une recherche que le thème.

Dans le cadre de cette thèse, nous avons proposé de modéliser à partir d'ontologies deux aspects du contexte : le thème abordé dans la collection sur laquelle l'utilisateur effectue sa recherche mais aussi la tâche de recherche que l'utilisateur choisit pour combler son besoin. La

[1] http://www.educnet.education.fr/cdi/ress/logdoc.htm
[2] http://fr.yahoo.com/

prise en compte de ce dernier aspect permet de cibler les informations du corpus qui intéressent l'utilisateur dans la tâche de recherche qu'il accomplit. L'intérêt des ontologies est que ces deux aspects du contexte sont précisés par rapport au domaine de connaissance qui leur est associé. Notre modèle intègre ainsi ces deux aspects modélisés au travers d'ontologies. Une des originalités importantes de notre approche est que les deux aspects sont liés à partir d'éléments communs aux deux ontologies. Ce lien est modélisé par la possibilité de spécifier des éléments du thème parmi les informations intéressant l'utilisateur dans sa tâche. L'intégration du modèle dans le processus de RI est réalisée à deux niveaux en préservant le lien. Les ontologies sont utilisées tout d'abord pour permettre l'indexation sémantique (au niveau des concepts) des granules de la collection. Les granules sont alors situés par rapport aux deux aspects de la modélisation du contexte. Les ontologies servent également de support d'accès à l'information. Par leur navigation, l'utilisateur a accès aux informations du corpus extraites à partir des deux ontologies. L'originalité du mécanisme de navigation que nous avons défini permet à l'utilisateur d'une part d'avoir une vue globale sur la collection par l'accès aux concepts des ontologies et, d'autre part, d'accéder aux informations élaborées par la visualisation des instances du modèle associé à l'ontologie de tâche. L'accès fait donc intervenir deux niveaux d'abstraction. De plus, à ces deux niveaux, le mécanisme de navigation distingue la connaissance relative à la tâche et celle relative au thème, tout en représentant le lien entre ces deux aspects.

Les ontologies que nous considérons dans notre modèle sont des ontologies légères. Ces ontologies sont composées d'un lexique, de concepts et de relations taxonomiques et associatives entre concepts. Bien qu'elles soient d'un niveau formel moindre que les ontologies lourdes, ces ontologies permettent, du point de vue de la RI, de mettre en place des mécanismes de recherche élaborés. Le lexique qu'elles spécifient est utilisé pour détecter les concepts référencés dans les documents et la structuration des concepts sert à la pondération de l'importance des concepts dans les documents. L'utilisation d'ontologies lourdes a été envisagée, mais leur intégration dans les SRI pose de nombreux problèmes. Tout d'abord, leur élaboration est très coûteuse car elles sont peu concevables à partir de procédés semi-automatiques. De plus, lorsque ces ontologies existent, elles sont rarement réutilisables pour une autre tâche que celle pour laquelle elles ont été élaborées. La connaissance qu'elles formalisent est en effet spécifique à la tâche initiale. Leur intégration dans un processus de RI est donc délicate. Le niveau formel est difficilement exploitable sur de grosses collections car leur utilisation pour la détection automatique des concepts qu'elles spécifient est longue à mettre en place. L'utilisation d'ontologies légères à forte composante lexicale, telle que nous la proposons, est donc une contribution importante dans l'intégration de connaissance dans les SRI.

La question de la réutilisabilité de ce type d'ontologies a été abordée dans cette thèse. Nous avons proposé une méthodologie et des méthodes visant à évaluer l'adéquation entre une ontologie légère de domaine et un corpus. Ceci permet de choisir, pour un processus de RI, une ontologie qui spécifie effectivement la connaissance abordée dans la collection considérée. L'originalité de notre méthodologie est qu'elle décompose les différentes étapes de l'analyse de l'adéquation, à savoir une pré-analyse des caractéristiques des deux ressources, l'analyse de leur lexique et l'analyse des relations entre concepts. Cette décomposition permet d'évaluer indépendamment les différents critères d'adéquation, ce qui n'est pas le cas dans les travaux de la littérature. L'intérêt des méthodes que nous proposons pour mettre en œuvre cette méthodologie repose, d'une part, sur l'extraction des termes représentatifs du corpus pour l'analyse lexicale de l'adéquation et, d'autre part, sur la prise en compte des relations entre concepts de l'ontologie qu'elles soient taxonomiques ou associatives lors de l'analyse conceptuelle, alors que les relations associatives, porteuses de connaissances, ne sont pas considérées dans les travaux de la littérature. Dans le cas où l'ontologie de domaine n'existe pas mais qu'une ressource lexicale telle qu'un thésaurus existe, nous avons proposé une méthode visant à élaborer une ontologie légère de

domaine à partir de la transformation semi-automatique du thésaurus. Une des originalités de notre approche est que la transformation repose sur l'analyse d'un corpus de référence pour expliciter et mettre à jour la connaissance contenue dans le thésaurus. De plus, notre méthode présente l'avantage de limiter l'intervention d'un expert dans cette transformation par la proposition d'éléments à chaque étape (labels de concepts, concepts, relations entre concepts).

Un prototype a aussi été développé pour mettre en œuvre nos propositions. Il repose sur de nombreuses fonctionnalités permettant la visualisation d'ontologies volumineuses formalisées en OWL. Il permet leur navigation par la sélection des concepts à visualiser et l'affichage des contextes des concepts à partir de leur contexte dans l'ontologie (labels, relations directes). Il met également en place un mécanisme de stockage des annotations des documents à partir des deux ontologies par l'implantation d'une base de données permettant un accès rapide aux informations du corpus élaborées à partir des deux ontologies. Les technologies de la conception d'une ontologie que nous avons développées ont été évaluées dans le cadre d'application de l'astronomie. Les astronomes qui ont évalué les résultats que nous avons obtenus, ont montré leur pertinence. Dans ce domaine, le thésaurus IAU construit en 1995 fait référence et les acteurs du domaine souhaiteraient l'utiliser plus largement. Le fait qu'il ne soit pas à jour est un frein à son utilisation. Notre contribution a amené un regain d'intérêt qui se concrétisera par une présentation dans le cadre du projet de l'Observatoire Virtuel International[3] et au congrès de l'« International Astronomical Union » (IAU)[4].

De nombreuses perspectives s'offrent à la suite de nos travaux. La première d'entre elles est d'étendre les expérimentations afin d'analyser plus en détail l'impact de nos différentes propositions. Dans un premier temps, de nouvelles expérimentations pourraient être réalisées dans le cadre applicatif considéré dans nos travaux. L'évaluation de la mesure de proximité entre concepts pourrait être étendue à d'autres couples de concepts de l'ontologie légère de l'astronomie. Les vingt couples de concepts que nous avons choisis sont certes représentatifs de l'ontologie, mais ne correspondent qu'à un faible pourcentage de ses concepts. La pertinence des concepts associés à chaque document par le mécanisme d'indexation sémantique que nous proposons pourrait être testée sur une collection plus importante. Nous avons choisi de l'évaluer sur dix documents sélectionnés avec précaution mais nous pourrions étendre l'évaluation. Dans un deuxième temps, des expérimentations pourraient être menées dans un autre cadre. La méthode de transformation d'un thésaurus en ontologie légère pourrait être appliquée à la transformation du thésaurus MOBIS utilisé dans le domaine de l'Education Nationale. Le prototype que nous avons développé pourrait ensuite être testé à partir de l'ontologie du thème ainsi créée et d'une ontologie liée à la tâche d'apprentissage pédagogique. L'accès à l'information réalisé pourrait être analysé dans ce contexte. Un sujet de DEA a d'ailleurs été proposé dans cette perspective.

L'amélioration du stockage des concepts des ontologies dans le prototype est aussi envisagée. Nous avons optimisé la gestion des instances relatives aux concepts de l'ontologie de tâche afin de faciliter leur accès dans le processus de restitution des informations à l'utilisateur. Nous pensons également stocker les concepts de façon persistante dans une base de données pour rendre plus rapide le chargement de l'ontologie dans le prototype.

Une autre perspective serait d'intégrer d'autres aspects du contexte à notre modèle. Une intégration qui est facilement envisageable est la prise en compte des préférences de l'utilisateur au niveau de l'accès aux documents. Ces préférences pourraient être incorporées à un profil et contiendraient le choix des relations à visualiser, la présentation des documents,… Un autre aspect serait d'ajouter des ontologies représentant la connaissance de l'utilisateur a priori sur la tâche et sur le thème. Ces ontologies pourraient être construites à partir d'informations que

[3] http://www.ivoa.net/
[4] http://www.astronomy2006.com/

l'utilisateur fournirait au préalable (à travers, par exemple, la sélection des concepts des deux ontologies qu'il maîtrise) et seraient mises à jour au fur à mesure qu'il consulte de nouveaux documents. Ces ontologies seraient utilisées pour filtrer l'information restituée à l'utilisateur et cibler les éléments aidant l'utilisateur à approfondir ou à élargir sa connaissance sur le domaine.

Une autre perspective serait d'étudier l'intégration d'ontologies lourdes pour lesquelles des axiomes seraient spécifiés. La prise en compte des axiomes permettrait la mise en place d'un mécanisme pour détecter les concepts non seulement à partir de leurs labels mais aussi à partir d'inférences sur les différents éléments qui identifient un concept (concepts équivalents, concepts dont il est l'union, valeurs pour une certaine relation,…). Le mécanisme d'indexation sémantique pourrait aussi être enrichi à partir d'inférences permettant une meilleure pondération des concepts dans les documents comme, par exemple, l'indexation à partir d'un concept dont les labels ne sont pas explicités dans le document mais qui peut être inféré des concepts détectés. Cependant, comme nous l'avons précisé précédemment, l'intégration de telles ontologies nécessite de nombreuses réflexions sur la conception ou la réutilisation de telles ressources pour la RI. Notre thèse représente donc une contribution à l'intégration de connaissances en RI. Une seconde étape sera de prendre en compte les ontologies lourdes.

Références

[Agirre 2000] 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.

[Alfonseca 2002] E. Alfonseca, S. Manandhar, Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures, EKAW-2002, Lecture Notes in Artificial Intelligence 2473, Springer Verlag, 2002.

[Allan 2003a] J. Allan, Hard Track Overview in TREC 2003: High Accuracy Retrieval from Documents, Text Retrieval Conference, http://trec.nist.gov/pubs/trec12/papers/HARD.OVERVIEW.pdf, 2003.

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

[Amous 2001] I. Amous, C. Chrisment, F. Sèdes, Extending the OOHDM methodology by eliciting metadata and generic structure, In Actes de la Conférence DATAKON, pp 249-256, 2001.

[Andreasen 2003] T. Andreasen, H. Bulskov, R. Knappe, Similarity for Conceptual Querying, In Proceedings for the 18th International Symposium on Computer and Information Sciences, pp 268-275, 2003.

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

[Arpirez 2003] J. Arpirez, O. Cororcho, M. Fernandez-Lopez, A. Gómez-Pérez, WebODE in a nutshell, AI Magazine, 24(3), pp 37-48, 2003.

[Assadi 1999] H Assadi, Construction of a regional ontology from text and its use within a documentary system. In Proceedings of the 2nd Formal Ontology in Information Systems Conference, N. Guarino (Ed.), pp 236–249, 1999.

[Assticcot 2003] Rapport de l'action spécifique ASSTICCOT, Action Spécifique STIC « Corpus et Terminologie » (AS 34), Rattachée au RTP-DOC (RTP 33), Rapport internet IRIT/2003-23-R, 2003.

[Aussenac 2000a] N. Aussenac-Gilles, B. Biébow, S. Szulman, Modélisation du domaine par une méthode fondée sur l'analyse de corpus, In Actes de la Conférence en Ingénierie des Connaissances (IC'2000), pp 93-103, 2000.

[Aussenac 2000b] N. Aussenac-Gilles, B. Biébow, S. Szulman, Revisiting ontology design : a method based on corpus analysis, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), R Dieng, O. Corby (Eds.), pp 172-188, 2000.

[Aussenac 2002] N. Aussenac-Gilles, B. Biebow, S. Szulman, Modelling the travelling domain from a NLP description with Terminae. In Proceedings of the 13th European Knowledge Acquisition Workshop (EKAW'02), 2002

[Aussenac 2004] N. Aussenac-Gilles, J. Mothe, Ontologies as Background Knowledge to Explore Document Collections, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), pp 129-142, 2004.

[Baader 1991] F. Baader, B. Hollunder, A Terminological Knowledge Representation System with Complete Inference Algorithms, In Proceedings of the Workshop on Processing Declarative Knowledge, 1991.

[Bachimont 1996] B. Bachimont, Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser, Thèse d'épistémologie, Ecole Polytechnique, Paris, 1996.

[Bachimont 1999] B. Bachimont, L'intelligence artificielle comme écriture dynamique : de la raison graphique à la raison computationnelle, Grasset, Paris, 1999.

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

[Bachimont 2004] B. Bachimont, Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle, Mémoire d'Habilitation à Diriger des Recherches, Université de Technologie de Compiègne, 2004.

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

[Banerjee 2002] S. Banerjee, T. Pedersen, An adapted Lesk algorithm for word sense disambiguation using Word-Net, In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, 2002.

[Baziz 2005] M. Baziz, M. Boughanem, N. Aussenac-Gilles, C. Chrisment. Semantic Cores for Representing Documents in IR, In Proceedings of the 20th ACM Symposium on Applied Computing, pp. 1020-1026, ACM Press ISBN: 1-58113-964-0, 2005

[Bechhofer 2001] S. Bechhofer, I. Horrocks, C. Goble, R. Stevens, OilEd: a Reasonable Ontology Editor for the Semantic Web, In Proceedings of the Joint German/Austrian Conference on Artificial Intelligence (KI'2001), volume 2174, pp 396–408, Springer-Verlag LNAI, 2001.

[Bechhofer 2003] S. Bechhofer, P. Lord, R. Volz, Cooking the Semantic Web with the OWLAPI, In Proceedings of the 2nd International Semantic Web Conference, 2003.

[Belkin 2004] N.J. Belkin, G. Muresan, X.M. Zhang, Using User's Context for IR Personalization, In Proceedings of the ACM/SIGIR Workshop on Information Retrieval in Context, 2004.

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

[Berners-Lee 2001] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American, pp 28–37, May 2001.

[Bernstein 2005] A. Bernstein, E. Kaufmann, C. Buerki, M. Klein, How Similar Is It? Towards Personalized Similarity Measures in Ontologies, In Proceedings of the 7 Internationale Tagung Wirtschaftsinformatik, pp 1347-1366, 2005.

[Borst 1997] P. Borst, Construction of Engineering Ontologies for Knowledge Sharing and Reuse, Ph.D Dissertation, Tweente University, 1997.

[Bourigault 1996] D. Bourigault, LEXTER, a Natural Language Processing Tool for Terminology Extraction, In Proceedings of 7th EURALEX International Congress, 1996.

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

[Bourigault 2002a] D. Bourigault, G. Lame, Analyse distributionnelle et structuration de terminologie documentaire du droit, Journal TAL, 43-1, 2002.

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

[Bourigault 2003] D. Bourigault, N. Aussenac-Gilles, Construction d'ontologies à partir de textes, In Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN), pp 27-50, 2003.

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

[Brachman 1977] R.J. Brachman, What's in a concept : structured foundation for semantic networks, International Journal of Man-Machine Studies 9, pp 127-152, 1977.

[Brachman 1985] R.J. Brachman, J. Schmolze, An overview of the KL-One knowledge representation system, Cognitive Science, 9(2), pp 171- 216, 1985.

[Bradley 2001] N. Bradley, The {XML} Companion, Addison-Wesley Professional Publisher, 2001.

[Brewster 2004] C. Brewster, H. Alani, S. Dasmahapatra, Y. Wilks, Data driven ontology evaluation, In Proceedings of 4th International Conference on Language Resources and Evaluation, 2004.

[Brini 2005] A. Brini, M. Boughanen, D. Dubois, A Model for Information Retrieval based on Possibilistic Networks, In Proceedings of the 12th Symposium on String Processing and Information Retrieval (SPIRE), à paraître, 2005.

[Bruandet 1983] M.F. Bruandet, Y. Chiaramella, D. Kerkouba, Méthodes empiriques de construction de thésaurus: expérimentation, Revue de CID, Janvier-Mars, 1983.

[Buckley 2000] C. Buckley, J. Walz. SMART in TREC 8. In E. M. Voorhees (Ed), Special issue: The sixth Text Retrieval Conference (TREC-6), Information Processing and Management, 36(1), January 2000.

[Budanitsky 2001] A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, In Proceedings of the Workshop on WordNet and Other Lexical Resources, ACL, 2001.

[Cardoner 2004] L. Cardoner, Rapport de stage de 3ème année d'IUP, Université Paul Sabatier, Toulouse, 2004.

[Caropreso 2000] M-F. Caropreso, S. Matwin, F. Sebastiani, A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization, In Text Databases and Document Management: Theory and Practice, A.G. Chin (Ed.), Idea Group Publishing, Hershey, US, pp 78-102, 2000.

[Charlet 2000] J. Charlet, G. Kassel, M. Zacklad, D. Borigault, Ingénierie des connaissances : recherches et perspectives, In Ingénierie des connaissances, Évolutions récentes et nouveaux défis, Eyrolles, Paris, pp 1-22, 2000.

[Charlet 2002] J. Charlet, L'ingénierie des connaissances, développements, résultats et perspectives pour la gestion des connaissances médicales, Mémoire d'Habilitation à Diriger des Recherches, Université Pierre et Marie Curie, Paris, 2002.

[Chevalier 2002] M. Chevalier, Interface adaptative pour l'aide à la recherche d'information sur le web, Thèse de doctorat, Université Paul Sabatier, Toulouse, 2002.

[Chinchor 1998] N. Chinchor, P. Robinson, Hub-4 Named Entity Task Definition (version 3.5), In Proceedings of the MUC-7, 1998.

[Chrisment 2006] C. Chrisment, B. Dousset, T. Dkaki, S. Karouach, J. Mothe, Combining Mining and Visualization Tools to Discover the Geographic Structure of a Domain, Computers, Environment and Urban Systems Journal, à paraître.

[Condamines 2005] A. Condamines, Sémantique et Corpus, Hermès Science Publications, ISBN 2-7462-1055-X, 2005.

[Corcho 2004] Ó. Corcho, A. Gómez-Pérez, R. González-Cabero, M.C. Suárez-Figueroa, ODEVAL: A Tool for Evaluating RDF(S), DAML+OIL and OWL Concept Taxonomies, in Proceedings of the 1st International Conference on Artificial Intelligence Applications and Innovations, pp 369-382, 2004.

[Cucchiarelli 2004] R. Cucchiarelli, R. Navigli, F. Neri, P. Velardi, Extending and Enriching WordNet with OntoLearn, In Proceedings of the 2nd Global WordNet Conference, 2004.

[Cunningham 2002] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.

[Dachlet 1990] R. Dachlet, Etat de l'Art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information, Rapport de recherche de l'INRIA-Rocquencourt, Avril 1990, Projet : PSYCHO-ERGO - 32 p.[On-line]- < URL : http://www.inria.fr/rrrt/rr-1201.html >, 1990.

[Daille,1996] B. Daille, Study and implementation of combined techniques for automatic extraction of terminology, In The Balancing Act: Combining Symbolic and Statistical, J. Klavans, P. Resnik (Eds.), MIT Press, 1996.

[David 1990] S. David, P. Plante, Termino version 1.0, Report, Centre d'Analyse de Textes par Ordinateur, Université du Québec, 1990.

[Davis 1979] R. Davis, Interactive transfer of expertise, Artificial Intelligence, 12(2):121Ð157, 1979.

[Davis 1993] R. Davis, H. Sorbe, P. Szolovits, What is a Knowledge Representation?, AI Magazine. Spring, pp 17-33, 1993.

[Decker 2000] S. Decker, M. Erdmann, D. Fensel, I. Horrocks, M. Klein, F. van Harmelen, Oil in a nutshell, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), 2000.

[Denjean 1989] P. Denjean, Interrogation d'un systeme videotex : l'indexation automatique des textes, Thèse de doctorat, Université Paul Sabatier, Toulouse, 1989.

[Dervin 1992] B. Dervin, From the mind's eye of the user: the sense-making qualitative-quantitative methodology, In Qualitative Research in Information Management, J.Glazier, R. Powell (Eds.), Englewood, Libraries Unlimited, 1992.

[Desmontils 2002] E. Desmontils, C. Jaquin, Indexing a Web site with a terminology oriented ontology, The Emerging Semantic Web, I. Cruz S. Decker, J. Euzenat, D.L. McGuinness (Eds.), IOS Press, ISBN 1-58603-255-0, pp 181-197, 2002.

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

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

[Domingue 1998] J. Domingue, Tadzebao and WebOnto: Discussing, Browsing and Editing Ontologies on the Web, In Proceedings of the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'98), 1998.

[Dousset 2003] B. Dousset, Intégration de méthodes interactives de découverte de connaissances pour la veille stratégique, Mémoire d'Habilitation à Diriger des Recherches, Université Paul Sabatier, Toulouse, 2003.

[Dumais 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6), pp 391-407, 1990.

[Dumais 1994] S. Dumais, Latent Semantic Indexing (LSI) and TREC-2, In D. Harman (Ed.), The Second Text Retrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215, pp 105-116, 1994.

[Dumais 1995] S. Dumais, Using LSI for information filtering: TREC-3 experiments, In D. Harman (Ed.), The Third Text Retrieval Conference (TREC3), National Institute of Standards and Technology Special Publication, 1995.

[Dumais 2000] S. Dumais, H. Chen, Hierarchical classification of Web content, In Proceedings of the 23rd ACM Conference on Research and Development in Information Retrieval (SIGIR'00), pp 256-263, 2000.

[Dumais 2003] S. Dumais, E. Cutrel, J. Cadiz, G. Jancke, R. Sarin, D. Robbins, Stuff I've Seen: A system for personal information retrieval and re-use, In Proceedings of the 26th ACM Conference on Research and Development in Information Retrieval (SIGIR'03), 2003.

[Ehrig 2005] M. Ehrig, P. Haase, N. Stojanovic, M. Hefke, Similarity for Ontologies - A Comprehensive Framework, In Proceedings of the 13th European Conference on Information Systems, 2005.

[Englmeier 2003] K. Englmeier, J. Mothe, IRAIA: A portal technology with a semantic layer coordinating multimedia retrieval and cross-owner content building, In Proceedings of the International Conference on Cross Media Service Delivery, Cross-Media Service Delivery Series, The International Series in Engineering and Computer Science, V. 740, pp 181-192, 2003.

[Erdmann 2000] M. Erdmann, A. Maedche, H. Schnurr, S. Staab, From manual to semi-automatic semantic annotation: About ontology-based text annotation tools, In Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, P. Buitelaar, K. Hasida (Eds.) 2000.

[Euzenat 2002] J. Euzenat, Eight questions about semantic Web annotations, IEEE Intelligent systems 17(2), pp 55-62, 2002.

[Faatz 2002] A. Faatz, R. Steinmetz, Ontology enrichment with texts from the WWW, In Proceedings of the 2nd Semantic Web Mining Workshop at ECMLI/PKDD, 2002.

[Fallside, 2001] D.C. Fallside, XMLSchema, World Wide Web Consortium (W3C), W3C Recommendation, http://www.w3.org/XML/Schema, 2001.

[Farquhar 1997] A. Farquhar, R. Fikes, J. Rice, The Ontolingua Server: a tool for collaborative ontology construction, International Journal of Human-Computer Studies, 46(6), pp 707–727, 1997.

[Faure 1998] D. Faure, C. Nedellec, A corpus-based conceptual clustering method for verb frames and ontology acquisition, In Proceedings of the LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, 1998.

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

[Fikes 1985] R. Fikes, T. Kehler, The Role of Frame-Based Representation in Reasoning, Communications of the ACM (CACM), 28(9), pp 904-920, 1985.

[Fischer 1998] D. H. Fischer, From Thesauri towards Ontologies?, In Structures and Relations in Knowledge Organization : Proceedings of the 5th International ISKO Conference, W.M. Hadi, J. Maniez, S. Pollitt (Eds.),Würzburg: Ergon, pp. 18-30, 1998.

[Fortier 2004] J.Y. Fortier, G. Kassel, Présentation "sur mesure" d'informations ; une approche appliquée aux mémoires organisationnelles, In Revue des Sciences et Technologies de l'Information : RSTI-RIA, numéro spécial sur les informations sur mesure, pp 515-547, 2004.

[Foskett 1977] D.J. Foskett, Thesaurus, Reproduced in Readings in Information Retrieval, P. Willett, K Sparck-Jones (Eds.), pp 111-134, 1977.

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

[Frakes 1992] W.B. Frakes, R Baeza Yates (Eds.), Information Retrieval Data Structures and Algorithms, Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[Freund 2005a] L. Freund, E.G. Toms, Using contextual factors to match intent, In Proceedings of the ACM SIGIR Workshop on Information Retrieval in Context (IriX), pp 16-21, 2005.

[Freund 2005b] L. Freund, E.G. Toms, Contextual search: from information behaviour to information retrieval, In Proceedings of the Annual Conference of the Canadian Association for Information Science, 2005.

[Furst 2004], F. Furst, Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation, Thèse de doctorat, Université de Nantes, 2004.

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

[Genesereth 1994] M.R. Genesereth, R.E. Fikes, Knowledge interchange format version 3.0 reference manual, http://logic.stanford.edu/kif/Hypertext/kif-manual.html, 1994.

[Gomez Perez 2000] A. Gómez-Pérez, A. Moreno, J. Pazos, A. Sierra-Alonso, Knowledge Maps: An essential technique for conceptualisation, In Data & Knowledge Engineering, 33(2), pp 169-190, 2000.

[Gomez Perez 1999] A. Gómez-Pérez, Evaluation of taxonomic knowledge in ontologies and knowledge bases, In Proceedings of the 12th Knowledge Acquisition for Knowledge-Based Systems Workshop, 1999.

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

[Gonzalo 1998] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrán, Indexing with WordNet synsets can improve text retrieval, In Proceedings of the COLING/ACL Workshop on Usage of WordNet for Natural Language Processing, 1998.

[Grefenstette 1992] G. Grefenstette, Use of syntactic context to produce term association lists for text retrieval, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), pp 89-97, 1992.

[Gruber 1993] T.R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition, 5 (2), pp 199-220, 1993.

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

[Gruninger 1995b] M. Grunninger, M. Fox, Methodology for the design and evaluation of ontologies, In Proceedings of the IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[Guarino 1994] N. Guarino, M. Carrara, P. Giaretta, Formalizing ontological commitments, In Proceedings of the AAAI conference, 1994.

[Guarino 1998a] N. Guarino, Some ontological principles for designing upper level lexical resources, In Proceedings of the 1st International Conference on Language Resources and Evaluation, 1998.

[Guarino 1998b] N. Guarino, Formal Ontology and Information Systems, In Formal Ontology in Information Systems, N Guarino (Ed.), IOS Press. pp 3-15, 1998.

[Guarino 1999] N. Guarino, C. Masolo, G.Vetere, OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, 14 (3), pp 70-80, 1999.

[Guarino 2000] N. Guarino et C. Welty, Identity, Unity, and Individuality: Towards a Formal Toolkit forOntological Analysis, In Proceedings of the European Conference on Artificial Intelligence (ECAI), pp 219-223, 2000.

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

[Guarino 2002] N. Guarino and C. Welty, Evaluating Ontological Decisions with OntoClean, In Communication of the ACM, 45(2), pp 61-65, 2002.

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

[Guo 2004] Y. Guo, Z. Pan, J. Heflin, An Evaluation of Knowledge Base Systems for Large OWL Datasets, In Proceedings of the International Semantic Web Conference, pp 274-288, 2004.

[Haav 2001] H.M. Haav, T.L. Lubi, A Survey of Concept-based Information Retrieval Tools on the Web, In Proceedings of the 5th East-European Conference ADBIS, Vol 2, pp 29-41, 2001.

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

[Harman 1992] D. Harman, The DARPA TIPSTER project, In SIGIR Forum, volume 26(2), pp 26-28, 1992.

[Harman 2004] D. Harman, C. Buckley, The NRRC reliable information access (RIA) workshop, In Proceedings of the 27th ACM SIGIR Conference on Research and Dvelopment in Information Retrieval, pp 528-529, 2004.

[Harper 1978] D.J. Harper, C.J. van Rijsbergen, An Evaluation of Feedback in Document Retrieval Using Co-Occurrence Data, Journal of Documentation, 34(3), pp 189-216, 1978.

[Harris 1968] Z. Harris, Mathematical Structures of Language, New-York, John Wiley & Sons, 1968

[Hawking 1999] D. Hawking, N. Craswell, P. Thistlewaite, D. Harman, Results and challenges in Web search evaluation, In Proceeding of the 8th International Conference on World Wide Web, pp 1321-1330, 1999.

[Hearst 1992] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, In Proceedings of the 14th International Conference on Computational Linguistics, 1992.

[Hearst 1997]. M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, In Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, pp 246-257, 1997.

[Heijst 1997] G. van Heijst, G. Schreiber, B. Wielinga, Using explicit ontologies for KBS development, International Journal of Human-Computer Studies, 42(2/3), pp 183-292, 1997.

[Hernandez 2003a] N. Hernandez, Etude de l'utilisation de syntagmes nominaux pour la catégorisation automatique de documents, In Actes de la conférence INFORSID, pp 53-68, 2003.

[Hernandez 2003b] N. Hernandez**,** Ontologies de tâche et de domaine pour l'aide à l'exploration d'une collection de documents, In Actes du colloque de l'EDIT (Ecole Doctorale Informatique et Télécommunications), 2003.

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

[Hernandez 2004b] N. Hernandez, N. Aussenac-Gilles, OntoExplo : Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine, In Actes de la VIème Journées de l'innovation, 2004.

[Hernandez 2004c] N. Hernandez, J. Mothe, Ontologies pour l'aide à l'exploration d'une collection de documents, In Actes de la conférence Veille Stratégique Scientifique & Technologique Systèmes d'information élaborée (VSST), 2004.

[Hernandez 2005a] N. Hernandez, Ontologies pour l'aide à l'exploration d'une collection de documents, In Revue des Sciences et Technologies de l'Information (RSTI), Série Ingénierie des systèmes d'information, 10 (1), pp 11-34, 2005.

[Hernandez 2005b] N. Hernandez, J. Mothe, S. Poulain, Accessing and mining scientific domains using ontologies: the OntoExplo System, Poster, In Proceedings of The 28th Annual International ACM SIGIR, pp 607-608, 2005.

[Hernandez 2006] N. Hernandez, J. Mothe, D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus, In Actes de la conférence Veille Stratégique Scientifique & Technologique (VSST), à paraître, 2006.

[Hersh 2004] W.R. Hersh et al., TREC 2004 Genomics Track Overview, http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf, 2004.

[Horrocks 2001] I. Horrocks, F. van Harmelen, P.F. Patel-Schneider, Reference description of the DAML+OIL (March2001) ontology markup language, http://www.daml.org/2001/03/reference.html, 2001.

[Hovy 1999] E.H. Hovy, C.Y.Lin, Automated Text Summarization in SUMMARIST, In Advances in Automatic Text Summarization, M. Maybury, I. Mani (Eds.), MIT Press, 1999.

[Jacquemin 1999] C. Jacquemin, E. Tzoukermann, NLP for term variant extraction_ A synergy of morphology lexicon and syntax, In Natural Language Information Retrieval, T. Strzalkowski (Ed.), pp 25-74, 1999.

[Jain 1999] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No 3, 1999.

[Jarvelin 1996] 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.

[Jarvelin 2004] K. Jarvelin, P. Ingwersen, Information seeking research needs extensions towards tasks and technology, Information Retrieval, 10 (1), pp 212, 2004.

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

[Johnson 2003] J.D Johnson, On contexts in information seeking, Journal of the American Society for Information Science, 39 (5), pp 735-760, 2003.

[Jones 2000] G.J.F. Jones, New Challenges for Cross-Language Information Retrieval: Multimedia Data and the User Experience, Lecture Notes In Computer Science; Vol. 2069, Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation, pp 72 - 82, 2000.

[Kahan 2001] J. Kahan, M. Koivunen, E. Prud'Hommeaux, R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In Proceedings of the 10th International World Wide Web Conference, pp 623-632, 2001.

[Karp 1999] R. Karp, V. Chaudhri, J. Thomer, Xol: An xml-based ontology exchange language. http://www.ai.sri.com/ pkarp/xol, 1999.

[Karvounarakis 2002] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, Rql: a declarative query language for rdf, In Proceedings of the 11th International World Wide Web Conference, pp 592-603, 2002.

[Kassel 1999] G. Kassel, S. Perpette, Cooperative ontology construction needs to carefully articulate terms, notions and objects, In Proceedings of the International Workshop on Ontology Engineering on the Global Information Infrastructure, 1999.

[Kassel 2002] G. Kassel, OntoSpec : une méthode de spécification semi-informelle d'ontologies, In Actes des 13èmes journées francophones d'Ingénierie des Connaissances (IC), pp 75-87, 2002.

[Kavalec 2004] M. Kavalec, A. Maedche, V. Svátek, Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning, In Proceedings of SOFSEM, pp 249-256, 2004.

[Kayser 1997] D. Kayser, La représentation des connaissances, Hermes, ISBN 2-86601-647-5, 1997.

[Khan 2002] L. Khan, F. Luo, Ontology Construction for Information Selection, In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, pp 122- 127, 2002.

[Kifer 1995] M. Kifer, G. Lausen, J. Wu, Logical Foundations of Object-Oriented and Frame-Based Languages, Journal of the ACM, 42(4), pp 741-843, 1995.

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

[Kullback 1951] S. Kullback, R.A. Leibler, On Information and Sufficiency, Annals of Mathematical Statistics **22**, 1951.

[Lame 2002] G. Lame, Construction d'ontologie à partir de texte, une ontologie du droit dédiée à la recherche d'information sur le Web, Thèse de doctorat, Ecole des Mines de Paris, 2002.

[Lamrous 1997] S.A. Lamrous, P, Trigano, Organisation des bases documentaires vers une exploitation optimale, Document numérique, 1(4), pp 459-481, 1997.

[Lassila 1999] O. Lassila, R. R. Swick, Resource description framework (rdf) model and syntax speci_cation w3c recommendation 22. février 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999.

[Lassila 2001] O. Lassila, D. McGuiness, The role of frame-based representation on the semantic Web, Rapport technique KSL-01-02, Knowledge Systems Laboratory, Stanford University, 2001.

[Lausen 2004] H. Lausen, M. Stollberg, R. Lara, Y. Ding, S.-K Han, D. Fensel, Semantic Web Portals: State of the Art Survey, Technical Report DERI-TR-2004-04-03, 2004

[Lawrie 2000] D. Lawrie, W. B. Croft, Discovering and Comparing Topic Hierarchies, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), pp 314-330, 2000.

[Le Moigno 2002] S. Le Moigno, J. Charlet, D. Bourigault, M.C. Jaulent, Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale, In Actes des 6 ^{es} Journées Ingénierie des Connaissances, pp 229-238, 2002.

[Leacock 1998] C. Leacock, M. Chodorow, Combining local context and Wordnet similarity for word sense identification, in WordNet: an electronic lexical database, C. Felbaum (Ed), Cambridge, MA, The MIT Press, pp 265-283, 1998.

[Lee 1995] J. Lee, G. Yost, P.W. Group, The PIF process interchange format and framework, Technical Report 180, MIT Center for Coordination Science, 1995.

[Lesk 1988] M. Lesk, "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems, In Proceedings of the 4th Annual Conference of the University of Waterloo Centre for the New OED, 1988.

[Liebowitz 1998] J. Liebowitz, T. Beckman, Knowledge Organizations: What Every Manager Should Know, St. Lucie Press, 1998.

[Lin 1998] D. Lin, An information-theoric definition of similarity, In Proceedings of the 15th international conference on Machine Learning, pp 296-304, 1998.

[Lin 2000] C.Y. Lin, E.H. Hovy, The Automated Acquisition of Topic Signatures for Text Summarization, In Proceedings of the COLING Conference, 2000.

[Lindeberg 1993] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods Inf Med, 32(4), pp 281-291, 1993. http://www.openclinical.org/medTermUmls.html

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

[Lovins 1968] J.B. Lovins, Development of a stemming algorithm, Mechanical translation and computational linguistics, Vo11. pp 22-31, 1968.

[Lozano-Tello 2004] A. Lozano-Tello, A. Gómez-Pérez, ONTOMETRIC: A Method to Choose the Appropriate Ontology, Journal of Database Management, 15(2), 2004.

[Luke 2000] S. Luke, J. Hein, Shoe 1.01 proposed specification, SHOE project, April 2000, http://www.cs.umd.edu/projects/plus/SHOE/spec.html, 2000.

[MacGregor 1991] R. MacGregor, Using a description classifier to enhance deductive inference, In Proceedings of the 7th IEEE Conference on AI Application, pp 93-97, 1991.

[Maedche 2000] A. Maedche, S. Staab, Mining ontologies from text, In Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, Springer Lecture Notes in Artificial Intelligence, (LNAI-1937), 2000

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

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

[Maedche 2002b] 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.), Spinger, 2002.

[Maedche 2003] A. Maedche et S. Staab, N. Stojanovic, R. Studer, Y. Sure, SEmantic portAL: The SEAL Approach, In Spinning the Semantic Web, D. Fensel, J.A. Hendler, H. Lieberman, W. Wahlster, (Eds.), MIT Press, Cambridge London, pp 317-359, 2003.

[Maedche 2004] A. Maedche, S. Staab, Ontology Learning, Handbook on Ontologies, S Staab, R. Stubers (Eds.), pp 173-190, 2004.

[Mandala 1999] 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, pp 191-197, 1999.

[Manning 1999] C.D. Manning, H. Schuetze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Massachusetts, 1999.

[McBride 2001] B. McBride. Jena: Implementing the rdf model and syntax specification, http://wwwuk.pl.hp.com/people/bwm/papers/20001221-paper/, Hewlett Packard Laboratories, 2001.

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

[McHale 1998] M. Mc Hale, A comparison of Wordnet and Roget's taxonomy for measuring semantic similarity, In Proceedings of the COLING/ACL Workshop on Usage of Wordnet in Natural Langage Processing Systems, pp 115-120, 1998.

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

[Miles 2003] A. J. Miles, N. Rogers, D. Beckett, Migrating thesauri to the semantic web, guidelines and case studies for generating RDF encodings of existing thesauri, SWAD Europe Thesaurus Activity, Delivrable 8.8, 2003 http://www.w3.org/2001/sw/Europe/reports/thes/8.8/

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

[Milks 2002] Y. Milks, Ontotherapy or how to stop worrying about what there is, In Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases, 2002.

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

[Miller 1991] G. Miller,W.G. Charles, Contextual Correlates of Semantic Similarity, Language and Cognitive Processes, 6(1), pp 1-28, 1991.

[Miller 1993] G. Miller, C. Leacock, R. Tengi, R.T. Bunker, A Semantic Concordance, In Proceedings of ARPA Workshop on Human Language Technology, pp 303-308, 1993.

[Miller 2002] L. Miller, A. Seaborne, A. Reggiori, Three implementations of squishql, a simple rdf query language, In Proceedings of the International Semantic Web Conference, pp 423-435, 2002.

[Minsky 1975] M. Minsky, A framework for representing knowledge, In Psychology of Computer Vision, P.H. Winston (Ed), pp 211-277, 1975.

[Mitra 1997] M. Mitra, C. Buckley, A. Singhal, C. Cardie, An analysis of Statistical and Syntactic Phrases, In Actes de la conférence Recherche d'Information Assistee par Ordinateur (RIAO), 1997.

[Moldovan 1999] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihal-cea, R. Goodrum, R. Girju, V. Rus, LASSO: A tool for surfing the answer net. In Proceedings of the 8th Text Retrieval Conference (TREU-8), 1999.

[Montaner 2003] M. Montaner, B. Lopez, J.L. De La Rosa, A taxonomy of recommender agents on the Internet. Artificial Intelligence Review, 19, pp 285-330, 2003.

[Morin 1999] E. Morin, Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus", In Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE'99), pp 268-278, 1999.

[Mothe 2003a] J. Mothe, C. Chrisment, B. Dousset, J. Alaux, DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets, Journal of the American Society for Information Science and Technology, Vol 54 (2), pp 650-659, 2003.

[Mothe 2003b] J. Mothe, G. Hubert, J. Augé, K. Englmeier, Catégorisation automatique de textes basée sur des hiérarchies de concepts, Journées Bases de Données Avancées, pp 69-87, 2003.

[Mothe 2004] J. Mothe, D. Egret, C. Chrisment, K. Englmeier, Knowledge discovery in bibliographic collections using concept hierarchies and visualisation tools Library and Information Services in Astronomy, LISA IV, pp 233-241, 2004.

[Murtagh 1998] F. Murtagh, Clustering and Classification, The Computer Journal, 41, p. 517, 1998.

[Nazarenko 2004] A. Nazarenko, Donner accès au contenu des documents textuels Acquisition de connaissances et analyse de corpus spécialisés, Mémoire d'Habilitation à Diriger des Recherches, Université Paris Nord, 2004.

[Nonaka 1995] I. Nonaka, H. Takeuchi, The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, 1995.

[Noy 2000] N. Noy, R.W. Ferguson, M.A. Musen, The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), 2000.

[Ok Koo 2003] 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.

[Paralic 2003] J.Paralic, I.Kostial, Ontology-based Information Retrieval, In Proceedings of the 14th International Conference on Information and Intelligent Systems, ISBN 953-6071-22-3, pp 23-28, 2003.

[Patwardhan 2003] S. Patwardhan, S. Banerjee, T. Pedersen, Using Measures of Semantic Relatedness for Word Sense Disambiguation, In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, pp 241-257, 2003.

[Pinto 2001] H.S. Pinto, J.P. Martins, A methodology for ontology integration, In Proceedings of the International Conference on Knowledge Capture, pp 131-138, ACM Press, 2001.

[Pohlmann 1997] R. Pohlmann, W. Kraaij, The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), L. Devroye, C. Chrisment (Ed.), pp 176-187, 1997.

[Poli 2002] R. Poli, Ontological methodology, International Journal of Human-Computer Studies, pp 639-664, 2002.

[Ponte, 1998] J. M. Ponte, W. B. Croft, A Language Modeling Approach to Information Retrieval, Research and Development in Information Retrieval, In Proceeding of the 21st International ACM-SIGIR conference on Research and Development in Information Retrieval, 1998.

[Porter 1980] M. Porter, An algorithm for suffix stripping Program, 14(3), pp 130-137, 1980.

[Porzel 2004] R. Porzel, R. Malaka, A Task-based Approach for Ontology Evaluation, In Proceedings of the ECAI Workshop on Ontology Learning and Population, pp 9-16, 2004.

[Quillian 1968] M.R. Quillian, Semantic memory, In Semantic Information Processing, pp 216-270, MIT press, Cambridge, 1968.

[Rada 1989] R. Rada, H. Mili, E. Bicknell, M.Blettner, Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man and Cybernetics, 19(1), pp 17-30, 1989.

[Rauber 2001] A. Rauber, A. Muller-Kogler, Integrating automatic genre analysis into digital libraries, In Proceedings of the 1st ACM-IEEE-CD Joint Conference on Digital Libray (JCDL), pp 1-10, 2001.

[Resnik 1995] P. Resnik, Using information content to evaluate similarity in a taxonomy, In Proceedings of the 14th joint conference in Artificial Intelligence, 1995.

[Resnik 1999] P. Resnik, Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural langage, Journal of Artificial Intelligence Research, volume 11, pp 95-130, 1999.

[Richarson 1995] R. Richardson, A.F. Smeaton, Using WordNet in a Knowledge-Based Approach to Information Retrieval, Working Paper, CA-0395, School of Computer Applications, Dublin City University, Ireland, 1995.

[Rieu 1999] D. Rieu, Ingénierie des systèmes d'information : bases de données, bases de connaissances, et méthodes de conception, Mémoire d'Habilitation à Diriger des Recherches, INP de Grénoble, 1999.

[Rijsbergen, 1979] C.J van Rijsbergen., Information Retrieval, Butterworths, London, 1979.

[Riloff 1996], E. Riloff, Automatically generating extraction patterns from untagged text, In Proceedings of the 13th National Conference on Artificial Intelligence, pp 1044-1049, 1996.

[Rivier 1990] A. Rivier, Construction des langages d'indexation : Aspects théoriques, Documentaliste, vol. 27 (6), pp 263-274, 1990.

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

[Roberston 1997] S.E. Robertson, S. Walker, M. Beaulieu, Okapi at TREC7, [TREC 7] pp 253-264, 1997.

[Roberston 2002] S. Roberston, I Soboroff, The TREC 2002 Filtering Track Report, http://trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.pdf, 2002.

[Rocchio 1971] J. Rocchio, Relevance Feedback in Information Retrieval, In The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton (Ed), pp. 313-323, 1971.

[Rocha 2004] C. Rocha, D. Schwabe, M.P. de Aragão, A Hybrid Approach for Searching in the Semantic Web, In Proceedings of the 13th International World Wide Web Conference, pp 374-383, 2004.

[Rothenburger 2002] B. Rothenburger, A Differential Approach for Knowledge Management, In Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, 2002.

[Sahami 1996] M. Sahami, Learning limited dependence Bayesian classifiers, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp 335-338, 1996.

[Saias 2003] J. Saias, P. Quaresma, A Methodology to Create Ontology-Based Information Retrieval Systems, In Proceedings of the EPIA Conference, pp 424-434, 2003.

[Salton 1971] G. Salton, The Smart Retrieval System, Prentice Hall, Englewood Cliffs, NJ, 1971.

[Salton 1983] G. Salton, M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[Salton 1990] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science, 44 (4), pp 288-297, 1990.

[Sanderson 1999] M. Sanderson, W.B. Croft, Deriving concept hierarchies from text, In Proceedings of the 22nd International ACM SIGIR Conference, pp 206-213, 1999.

[Sanderson 2000] M. Sanderson, Retrieving with good sense, In Information Retrieval Vol. 2 No. 1, pp 49-69, 2000.

[Savoy 1993] J. Savoy, Stemming of French Words Based on Grammatical Categories, Journal of the American Society for Information Science, 44(1) pp 1-9, 1993.

[Schmid 1994] H. Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, In Proceedings of the International Conference on New Methods in Language Processing, 1994.

[Schreiber, 1995] G. Schreiber, B. Wielinga, W. Jansweijer, The KACTUS view of the 'O' word, In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI'1995), 1995.

[Sebastiani 2002] F. Sebastiani, Machine Learning in Automated Text Categorisation, ACM Computing Surveys, vol 34, No. 1, pp 1-47, 2002.

[Sebastiani 2006] F. Sebastiani, Classification of text, automatic, In The Encyclopedia of Language and Linguistics, K. Brown (Ed.), Volume 14, 2nd Edition, Elsevier Science Publishers, Amsterdam, NL, 2006 (http://www.math.unipd.it/~fabseb60/Publications/ELL06.pdf).

[Seeling 2003] C. Seeling, A. Becks, Exploiting Metadata for Ontology-Based Visual Exploration of Weakly Structured Text Documents, In Proceedings of the 7th International Conference on Information Visualisation (IV03), IEEE Press, ISBN 0-7695-1988-1, pp 652-657, 2003.

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

[Shadbolt 1993] N. Shadbolt, E. Motta, A. Rouge, Constructing knowledge based systems. IEEE Software, 10(6), pp 34-38, 1993.

[Small 1982] S. Small, C. Rieger, Parsing and comprehending with word experts (a theory and its realisation), in Strategies for Natural Language Processing, W.G. Lehnert & M. H. Ringle (Eds.), pp 89-148, 1982.

[Smith 2004] M.K. Smith, C. Welty, D.L. McGuinness, OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-guide/

[Soboroff 2003] I. Soboroff, D. Harman, Overview of the TREC 2003 Novelty Track, http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf

[Soergel 1974] D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, Los Angeles, Melville Publ. Company, 1974.

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

[Sowa 1984] J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Publishing Company, USA, 1984.

[Spärck Jones 1996] K. Spärck Jones, Further reflections on TREC, Information Processing and Management, Vol 36 N 2000, pp 37-85, 1996.

[Spärck Jones 2003] K. Spärck Jones, Document Retrieval : Shallow Data, Deep Theories ; Historical Reflections, Potential Directions, In Proceedings of the 25th European Conference on IR Research, pp 1-11, 2003.

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

[Spiteri 1999] L. Spiteri, The essential elements of faceted thesauri, Cataloging & Classification Quarterly, 28(4), pp 31-52, 1999.

[Srikant 1995] R. Srikant, R. Agrawal, Mining generalized association rules, In Proceedings of the 21st Conference on Very Large DataBases (VLDB'95), pp 407-419, 1995.

[Staab 2000] S. Staab, A. Maedche, Axioms are objects too: Ontology engineering beyond the modeling of concepts and relations, Research report 399, Institute AIFB, Karlsruhe, 2000.

[Stevenson 2002] M. Stevenson, Combining disambiguation techniques to enrich an ontology, In Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-02) workshop on "Machine Learning and Natural Language Processing for Ontology Engineering", 2002.

[Stuckenschmidt 2004] H. Stuckenschmidt, F. van Harmelen, A. de Waard, T. Scerri, R. Bhogal, J. van Buel, I.Crowlesmith, C. Fluit, A. Kampman, J. Broekstra, E. van Mulligen, Exploring large document repositories with RDF technology: the DOPE project, Intelligent system, IEEE, Vol. 19, No. 3, pp 34- 40, 2004.

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

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

[Sugiua 2004] N.Sugiura, Y. Shigeta, N. Fukuta, N. Izumi, T. Yamaguchi,: Towards On-the-Fly Ontology Construction – Focusing on Ontology Quality Improvement. In Proceedings of the 1st European Semantic Web Symposium (ESWS), 2004.

[Sure 2002] Y. Sure, J. Angele, S. Staab, OntoEdit: Guiding Ontology Development by Methodology and Inferencing, In Proceedings of the Confederated International Conferences CoopIS, DOA and ODBASE 2002, volume 2519, pp 1205–1222, Springer-Verlag LNCS, 2002.

[Sure 2003] Y. Sure, Ó. Corcho, Evaluation of Ontology-based Tools, In Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools held at the 2nd International Semantic Web Conference, 2003.

[Taylor 1968] R.S. Taylor, Question negotiation and information seeking in libraries. College and Research Libraries 29, pp 178-194, 1968.

[Thieu 2004] M. Thieu, O. Steichen, Ch. Le Bozec, E. Zapletal, M.-Ch. Jaulent, Mesures de similarité pour l'aide au consensus en anatomie pathologique, In Proceedings of the 5th International Conference on Internet Computing, pp 225-236, 2004.

[Todirascu 2001] A. Todirascu, F. Rousselot, Ontologies for information retrieval, In Actes de la 8e conférence sur le Traitement Automatique des Langues Naturelles (TALN), 2001.

[Tudhope 2001] D. Tudhope, H. Alani, C. Jones, Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, 1-8(41), 2001.

[Turtle 1991] H.R. Turtle, Inference Networks for Document Retrieval, PhD Thesis, University of Massachusetts, 1991.

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

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

[Uschold 1998] M. Uschold, M. Healy, K.Williamson, P. Clark, S.Woods, Ontology Reuse and Application, In Proceedings of the International Conference on Formal Ontology and Information Systems, pp 179-192, 1998.

[Uschold 2003] M. Uschold, Where are the semantics in the semantic Web?, AI Magazine, 24(3), pp 25-36, ISSN:0738-4602, 2003.

[Vakkari 2003] P. Vakkari, Task-based information searching, Annual Review of Information Science and Technology, 37, pp 413-464, 2003.

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

[Van Assem 2004] M. van Assem, M. Menken, G. Schreiber, J. Wielemaker, B. Wielinga, A Method for Converting Thesauri to RDF/OWL, In Proceedings of the 3rd International Web Conference, 2004.

[Velardi 2001] P. Velardi, P. Fabriani, M. Missikoff, Using text processing techniques to automatically enrich a domain ontology, In Proceedings of the ACM Conference on Formal Ontologies and Information Systems, pp 270-284, 2002.

[Véronis 1989] J. Véronis, N. Ide, N. Wurbel, Extraction d'informations sémantiques dans les dictionnaires courants, In Actes du 7ème congrès Reconnaissance des Formes et Intelligence Artificielle, pp 1381-1395, 1989.

[Voorhees 1993] E.M. Voorhees, Using WordNet to disambiguate word sense for text retrieval, In Proceedings of the 13th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 171-180, 1993.

[Voorhees 1994] E.M. Voorhees, Query expansion using lexical-semantic relations, In Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 61-69, 1994.

[Voorhees 2004] E.M. Voorhees, D.M. Tice, The TREC-8 Question Answering Track Evaluation, http://trec.nist.gov/pubs/trec13/papers/QA.OVERVIEW.pdf, 2004.

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

[Woods 1975] W.A. Woods, What's in a link: Foundation for Semantic Networks, In Representation and Understanding; Studies in Cognitive Science, D.G. Bobrow, A. Collins (Eds.), Academic Press, pp 35-82, 1975.

[Woolf 1990] H. Woolf (Ed.), Websters New World Dictionary of the American Language, G. & C.Merrian, 1990.

[Wu 1994] Z. Wu, M. Palmer, Verb semantics and lexical selection, In Proceedings of the 32nd annual meeting of the Association for Computational Linguistics, pp 133-138, 1994.

[Xu 2000] J. Xu, W.B. Croft, Improving the Effectiveness of Information Retrieval with Local Context Analysis, ACM Transactions of Information Systems, 18(1), pp 79-112, 2000.

[Zamir 1998] O. Zamir O. Etzioni, Web Document Clustering : a Feasibility Demonstration**,** In Proceedings of the 21st ACM/SIGIR Conference on Research and Development in Information Retrieval, 1998.

[Zipf 1949] G. Zipf, Human Behaviour and the Principle of Least Effort, Addison-Wesley, 1949.

[Zhang 2004] S. Zhang, O. Bodenreider, Comparing Associative Relationships among Equivalent Concepts across Ontologies, In Proceedings of MEDINFO 2004, pp 459-464, 2004.

[Zweigenbaum 1993] P. Zweigenbaum et al., Linguistic and medical knowledge bases: An access system for medical records using natural language, Technical report, MENELAS: deliverable 9, AIM Project A2023, 1993.