Chapitre 1

# Ontologies pour la recherche d'information, importance de la dimension terminologique

## 1.1. Introduction

Depuis environ 10 ans, le contexte de la recherche d'information (RI) se modifie de manière accélérée et se trouve bousculé par l'intérêt dont témoignent l'intelligence artificielle et l'ingénierie des connaissances (IC) sur cette question. D'une part, avec le projet du Web Sémantique [BER 99], l'informatique en général (et particulièrement l'ingénierie des connaissances) affiche la volonté que les logiciels produits puissent aborder au niveau du sens, des idées et des contenus l'information présente sur Internet mais aussi dans les documents des communautés scientifiques et des professionnels. D'autre part, la numérisation accélérée des collections et la généralisation de la production de documents structurés à l'aide de XML rendent indispensables des techniques plus fines et pertinentes pour que les utilisateurs ne soient pas noyés sous les flots informationnels et puissent en tirer profit pour des objectifs précis. Or, parmi les moyens affichés par le projet du Web Sémantique [CHA 05], une des priorités est de disposer de modèles consensuels et structurés, définissant les notions-clés d'un domaine et permettant de raisonner sur ces connaissances : les ontologies.

Dans cet article, nous abordons la question de l'utilisation des ontologies pour la recherche d'information du point de vue de l'ingénierie des connaissances, de la recherche sur les ontologies, sur leur mode de construction et sur la nature de leur contenu. Nous défendons l'idée que les ontologies sont des représentations des connaissances d'autant plus pertinentes dans ce contexte qu'elles comportent une

---

Chapitre rédigé par Nathalie AUSSENAC-GILLES, Nathalie HERNANDEZ et Mustapha BAZIZ

dimension terminologique, et même qu'elles peuvent être rattachées à des éléments linguistiques comme des patrons d'extraction d'information. La richesse terminologique est en effet à la fois une conséquence de la construction des ontologies à partir de textes et un atout pour les utiliser dans une indexation conceptuelle. Enfin, les outils du Traitement Automatique du Langage (TAL) reposent sur des éléments linguistiques comme les patrons d'extraction de relations ou de concepts qui peuvent faciliter le processus d'annotation ou d'indexation. L'indexation revient alors à chercher des instances de concepts dans les textes à l'aide de ces patrons : elle permet à la fois de "peupler" l'ontologie de nouvelles instances de concepts et d'indexer les textes.

L'article s'organise en six parties. Nous présentons tout d'abord (partie 1.2) les différents types de ressources terminologiques et ontologiques qui sont appelées ontologies. Parmi les caractéristiques qui permettent de les comparer, nous montrons la place très variable qu'elles accordent aux termes du domaine. Par la suite, nous parlerons de *ressources termino-ontologiques* (RTO). . Nous développerons alors les atouts attendus des RTO pour la recherche d'information tels qu'ils sont formulés (partie 1.3) avant d'illustrer concrètement l'apport (et le coût) de ce type d'approche sur trois études de cas (partie 1.4). Pour finir (partie 1.5), nous insisterons sur l'importance de la dimension terminologique des RTO (et de son traitement) qui est un des facteurs déterminants dans ces approches, avant de conclure sur les nombreuses perspectives (pour le TAL, la RI et l'IC) qui permettraient d'améliorer l'exploitation sémantique des documents textuels.

## 1.2. Ontologies et ressources termino-ontologiques

### *1.2.1 Héritage pluridisciplinaire des ontologies*

Si la notion d'ontologie connaît aujourd'hui un succès exceptionnel, c'est certainement parce qu'elle est l'aboutissement de l'évolution de modèles de connaissances dans des domaines très différents. Les chercheurs du Web Sémantique insistent sur l'héritage philosophique (champ de l'Ontologie) et logique (ontologies formelles). Ce courant fait des concepts de l'ontologie de futurs prédicats logiques permettant de raisonner, essentiellement par classification. Ils situent également les ontologies dans l'héritage des recherches sur la représentation des connaissances, dans la lignée des réseaux sémantiques et des modèles conceptuels d'une part, et de leur représentation logique (travaux sur les langages de *frames* et les *logiques de description*). Ce point de vue renvoie aux motivations initiales du développement des ontologies en IC telles qu'elles sont formulées dans [GRU 91]. Les ontologies répondent à des besoins de formalisation, d'interopérabilité et de standardisation des modèles pour favoriser leur réutilisation,

faciliter leur maintenance et surtout mieux assurer les échanges de connaissances entre systèmes formels ou entre applications informatiques et utilisateurs [STA 04].

À ces éléments, le développement massif d'applications pour le web a soudain ajouté de nouveaux enjeux, à la fois techniques et économiques, ayant des conséquences sur la forme mais aussi le fond des modèles attendus. Les propositions d'architecture ou d'applications pour le futur web dit « web sémantique » font systématiquement appel aux ontologies : elles doivent fournir des représentations partagées utilisables par des agents logiciels, des bases de méta-données pour annoter ou indexer des documents ou encore assurer la mise à disposition de tous de bases de connaissances consensuelles.

Mais on peut situer aussi les ontologies dans une tradition terminologique, qui s'interroge depuis les années 40 sur les notions de termes et de concepts, sur l'articulation entre langue et connaissances, ou encore dans une tradition documentaire, dans la lignée des langages documentaires et des thésaurus. La gamme des produits à base terminologique nécessaires pour répondre aux besoins de la gestion documentaire s'élargit considérablement [BOU 00b]. À côté des bases de données terminologiques multilingues classiques, définies pour l'aide à la traduction, différents types de ressources terminologiques ou ontologiques (RTO) sont adaptés aux nouvelles applications de la terminologie en entreprise : glossaires et liste de termes pour les outils de communication interne et externe, thesaurus pour les systèmes d'indexation automatiques ou assistés, index hypertextuels pour les documentations techniques, terminologies de référence pour les systèmes d'aide à la rédaction, ontologies pour les mémoires d'entreprise, etc. [AUS 04b]. Plusieurs auteurs ont situé ces structures de données dans un continuum dont la dimension principale est le degré de formalisation (figure 1). Or les différences sont plus complexes et justifient de revenir sur ce que sont ces structures.

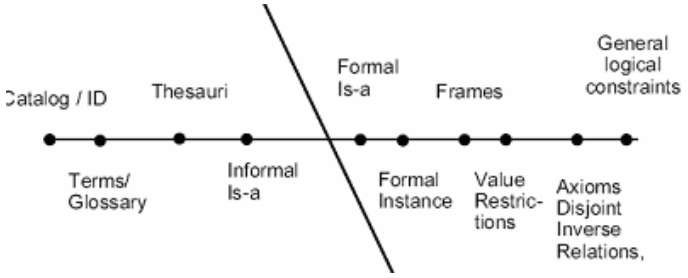


**Figure 1** *: Différentes ressources terminologique et ontologies selon leur degré de formalisation[LAS 01]*

### *1.2.2 Ressources utilisées en recherche d'information*

Dans la tradition des sciences de l'information et de la recherche documentaire, des ressources analogues servent à organiser des collections et à y retrouver des documents [AUS 03]. Le processus central est celui de l'*indexation***,** destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération. Un *index* est une table alphabétique des mots, des termes correspondant aux sujets traités, des noms cités dans un livre. Dans le domaine technique, l'index d'un document est aussi son analyse sommaire présentée sous forme de mots-clés, rubriques, etc. Pour constituer cette indexation par mots-clés, on peut puiser dans un *langage documentaire*, ensemble organisé de termes normalisés d'un domaine. Cette normalisation est au service d'une caractérisation du contenu des documents qui facilite une recherche ultérieure par une communauté d'utilisateurs. On distingue essentiellement, dans les langages documentaires, les classifications et les thesaurus. Une *classification* est la répartition systématique en classes, des termes désignant des êtres, choses ou notions ayant des caractères communs notamment afin d'en faciliter l'étude. Un *thesaurus* est un langage documentaire fondé sur une structuration hiérarchisée, alphabétique au premier niveau puis thématique, les termes normalisés étant reliés à des termes plus précis.

Avec l'informatisation de la recherche d'information, la notion d'index a évolué : un index construit automatiquement pour une recherche automatisée comporte une bien plus grande quantité d'éléments, qui ne sont plus forcément des termes, mais des chaînes de caractères (souvent associées à des valeurs numériques, des poids, afin d'accentuer leur pouvoir discriminant) construites à partir des mots des textes par troncature, lemmatisation, élimination de mots vides, etc. [MOT 00]. Parce qu'ils seront traités automatiquement, ces éléments n'ont plus besoin d'avoir du sens pour l'humain. Ils sont déterminés pour optimiser l'association entre une requête d'utilisateur et les documents correspondant le mieux à sa recherche. De ce fait, la richesse de l'index et l'ajustement des pondérations sont privilégiés par rapport à la normalisation des descripteurs.

Avec l'utilisation des ontologies pour une indexation sémantique, on revient à un index qui serait plus conforme à ceux des documentalistes : faire appel à une ressource normalisée pour caractériser un contenu informationnel. Pour comprendre les enjeux de ce basculement, revenons à la définition d'une ontologie en ingénierie des connaissances.

### *1.2.3 Ontologies en ingénierie des connaissances : définitions*

La notion d'ontologie a été redéfinie au gré des débats dont elle a fait l'objet. Les premières définitions présentent une ontologie comme une représentation formelle des connaissances [GRU 91], un « vocabulaire et des définitions des concepts d'un domaine » [USH 96]. La définition fondatrice de [GRU 93], a été actualisée dans [STU 98] sous la forme suivante : *An ontology is a formal, explicit specification of a shared conceptualisation.* Charlet en propose une définition complémentaire [CHA 02] : *Une ontologie est une spécification normalisée représentant les classes des objets reconnus comme existant dans un domaine. Construire une ontologie, c'est aussi décider d'une manière d'être et d'exister des objets de ce domaine.* Ainsi, une ontologie répond à des exigences complémentaires et symétriques : (i) en tant que spécification, elle définit une représentation formelle des connaissances permettant son exploitation par un ordinateur ; (ii) en tant que reflet d'un point de vue – partiel – sur un domaine, que l'on cherche le plus consensuel possible, elle fournit une sémantique qui doit permettre de relier la forme exploitable par la machine à sa signification pour les humains.

Concrètement, une ontologie modélise les connaissances d'un domaine sous forme d'un réseau de concepts normalisés et d'axiomes. Les concepts sont des classes génériques, définies par leurs relations sémantiques ou leurs propriétés (définition en intension par des conditions nécessaires et suffisantes) ou par la liste des instances relevant de cette classe (définition en extension). L'organisation des concepts est choisie de manière à favoriser leur classification : la structure d'une ontologie comporte donc systématiquement une hiérarchie de spécialisation des concepts et des relations définies par les concepts qu'elles relient (fig. 2). Selon les formalismes, les propriétés et les relations sémantiques entre concepts sont ou non héritées des classes vers leurs sous-classes.

Une structure d'ontologie est un quintuplet $O := \{C, R, H^C, rel, A^O\}$

C et R : ensembles disjoints des **concepts** et des **relations**

$H^C$ **hiérarchie** (taxonomie) de concepts : $H^C$ Í C x C , $H^C(C_1, C_2)$ signifie que $C_1$ est un sous-concept de $C_2$ (relation orientée)

Q Rel : **relation** rel: R ® C x C (définit des relations sémantiques non taxonomiques) avec 2 fonctions associées

dom : R ® C avec dom(R):= O1(rel(R))

range : R ® C avec range(R):= Õ2(rel(R)) co-domaine

rel(R) = $(C_1,C_2)$ s'écrit aussi $R(C_1,C_2)$

$A^O$ : ensemble **d'axiomes**, exprimés dans un langage logique adapté (logique de description, logique du 1er ordre)

**Figure 2** : *Définition de la structure d'ontologie dans [MAE 02a]*

On parle *d'ontologie légère* pour faire référence à une ontologie ne comprenant qu'une hiérarchie de concepts et les relations associées, *d'ontologie lourde* lorsqu'elle comporte aussi des axiomes. En effet, la définition d'axiome impose de disposer d'un langage logique adapté [GOM 04].

Le cadre de cette définition offre une grande liberté pour choisir et définir les concepts. Pour les tenants de l'ontologie formelle, l'ontologie présente des définitions consensuelles des concepts renvoyant à leur essence, aux concepts tels qu'ils sont échafaudés par les théoriciens d'un domaine, en dehors du contexte particulier pour lequel on s'y intéresse [GUA 00]. Une vue plus pragmatique considère que les concepts sont définis à travers la manière dont une communauté, scientifique ou technique, les manipule à travers le langage. Ces ontologies devant être intégrées au sein d'applications ciblées, l'application et la tâche à réaliser doivent être pris en compte. On parle d'ontologies régionales [BAC 04]. Il y a donc débat sur la généricité des concepts d'une ontologie, sur la prise en compte de leur utilisation dans leur définition, sur leur degré de formalisation, sur le domaine de couverture de l'ontologie (un domaine particulier versus les connaissances générales), mais aussi sur les principes appliqués pour organiser les concepts [BOU 04]. Des principes méthodologiques peuvent guider la manière de définir les concepts (quelles propriétés retenir) en indiquant comment les différencier les uns des autres par leurs propriétés (TERMINAE [AUS 04c], ARCHONTE [BAC 04]) ou en les situant par rapport à des classes génériques (ONTOSPEC [KAS 02], ONTOCLEAN [GUA 00]) ou des connaissances de sens commun. La légitimité des définitions de concepts peut renvoyer à l'expression des connaissances dans la langue (TERMINAE, ARCHONTE) ou encore aux méta-propriétés vérifiées par les concepts et les relations (ONTOCLEAN). Enfin, les ontologies couvrent des réalités différentes suivant leur utilisation, selon qu'elles soient destinées à être des connaissances partagées entre agents logiciels, des supports pour des systèmes interagissant avec l'utilisateur ou encore des ressources de méta-données pour indexer ou annoter des documents.

### *1.2.4 Composante lexicale et éléments linguistiques dans les ontologies*

Alors que les définitions « canoniques » des ontologies en IC les présentent parfois comme le vocabulaire d'un domaine, les langages de représentation font peu cas des termes associés aux concepts. Les réflexions de la terminologie en tant que discipline sur ses productions et sur l'articulation terme-concept éclaire pourtant la réflexion sur la notion de concept et leur statut par rapport aux termes et à leurs usages. Il nous paraît fondamental d'identifier en tant que telle la composante lexicale associée à une ontologie, par exemple selon la définition proposée sur la figure 3. Une ontologie à composante lexicale est alors un couple $(O, \mathcal{L})$ où $O$ est une ontologie et $\mathcal{L}$ son lexique.

Le lexique d'une structure d'ontologie $O := \{C, R, H^C, rel, A^O\}$ est un quadruplet $L := \{L^C, L^R, F, G\}$

$L^C$ et $L^R$: ensembles disjoints des **entrées lexicales** des concepts et des relations

$F, G$ : deux relations appelées **références :** $F$ pour les concepts, $G$ pour les relations

Pour L, *entrée lexicale de* $L^C$ : $F(L) = \{C$ *de* $C / (L,C)$ *est dans* $F\}$ *et* $F^{-1}(L) = \{L$ *de* $L / (L,C)$ *est dans* $F\}$

Idem pour $G$ *et* $G^{-1}$

**Figure 3** : *Définition du lexique d'une ontologie dans [MAE 02a]*

Pour étudier l'articulation entre connaissances et terminologie d'un domaine, des supports particulièrement intéressants sont les bases de connaissances terminologiques (BCT) [AUS 01]. Les BCT constituent un enrichissement significatif des terminologies traditionnelles sur papier car elles comportent une trace des informations conceptuelles relevées par le terminologue en identifiant les termes. Leur originalité est avant tout leur modèle de structuration des connaissances terminologiques. Ce modèle différencie un niveau linguistique d'un niveau conceptuel : on accède ainsi par les termes du domaine à une modélisation conceptuelle qui donne sens à ces termes [AUS 01]. La structure de cette composante, de type réseau sémantique, est proche de celle d'une ontologie [SZU 02]. Mais son contenu n'a pas d'ambition ontologique, ce réseau ne prétend ni normaliser ni standardiser ni fixer définitivement les définitions des concepts du domaine concerné. Au contraire, il rend compte des concepts tels qu'ils se dégagent de l'étude de la langue, en restituant une sémantique qui reste au plus près de l'usage terminologique. Les changements récents de la pratique terminologique, associés à un renouvellement théorique, remettent en effet en question les positions structuralistes [AUS 03].

Des réflexions interdisciplinaires ont conduit à répercuter sur les ontologies les résultats établis pour les BCT, en particulier pour la représentation des connaissances ainsi que l'intérêt de conserver, avec le modèle, les éléments de fouille de texte utiles à leur identification [SZU 04]. Nous entendons par là des patrons associés aux types de relations sémantiques (relation hiérarchique entre concepts et autre), des prédicats permettant de repérer des classes sémantiques, de calculer des synonymies, etc.

### *1.2.5 Construction d'ontologies à partir de textes*

Dès 1995, des méthodes de construction d'ontologie ont été proposées, définissant des canevas assez généraux, analogues à ceux du génie logiciel. Ces méthodes mettent l'accent sur la réutilisation d'ontologies existantes et sur l'exploitation de langages standard. Un panorama de ces méthodes est disponible

dans [GOM 04]. La définition de standards en matière de représentation des connaissances a été l'objet d'initiatives d'organismes comme la Darpa (langages KIF puis DAML) puis le W3C[1] avec l'avènement de OWL en 2000. L'utilisation de textes comme sources de connaissances n'est pas nouvelle en IC, mais la visée de définir des ontologies à partir de textes a renouvelé cette problématique. Ce regain s'explique en partie par des avancées du TAL, qui ont permis de dépouiller non plus des études statistiques des textes, mais les résultats de leur analyse selon des méthodes linguistiques. Cette approche n'a pas cessé de prendre de l'importance et de donner lieu à un nombre croissant de travaux. Historiquement, nous distinguerons trois périodes dont les contributions s'accumulent et sont toujours d'actualité.

De 1995 à 2000, les premières propositions sont avant tout méthodologiques, comme celles du groupe TIA[2]. L'innovation consiste à croiser l'expérience des terminologues à celle de la tradition des logiciens pour étudier les connaissances de domaines bien cernés : les méthodes sont supervisées pour laisser la place à l'interprétation humaine ; le statut des concepts est défini par rapport à ceux des termes en usage dans les textes ; des critères de normalisation (comme la différenciation dans la méthode Archonte [BAC 04]) guident l'organisation des concepts. Ces méthodes font appel à des logiciels de TAL existants, indépendants les uns des autres, comme les extracteurs de termes ou de classes sémantiques, les analyseurs syntaxiques robustes ou les concordanciers [BOU 02].

Progressivement (à partir de 1998), des logiciels spécialisés (combinant analyses statistiques et linguistiques) ont été développés pour des tâches propres à la construction d'ontologies : extracteurs de concepts (i.e. Syntex [BOU 02] ou SystemQuirk [AHM 95]) et extracteurs de relations sémantiques (i.e. Prométhée [MOR 99] et Caméléon [SEG 00]). Ces travaux visent toujours des domaines de spécialité. Leur succès conduit à de nouveaux développements pour intégrer différentes techniques et logiciels au sein de plates-formes. C'est ainsi qu'un module TAL a été ajouté à l'éditeur Protégé [BUI 04], un autre à WebODE [ARP 03], et que l'atelier KAON d'édition d'ontologie intègre Text-to-Onto [CIM 05] pour extraire des éléments d'ontologie des textes. Enfin, des techniques d'apprentissage contribuent à automatiser l'identification de classes sémantiques (ASIUM) ou à calculer des règles d'association entre concepts à partir de régularités d'usage entre termes [MAE 02b] [NAU 06].

Depuis 2003, on assiste à une véritable explosion de l'exploitation des textes, avec deux tendances fortes : la recherche systématique de l'automatisation et l'utilisation du web comme corpus. On parle alors « d'Ontology Learning and Population » pour renvoyer à l'identification de concepts et relations (learning) ou à

[1] http://www.W3C.org/
[2] http://www.loria.fr/TIA

celle d'instances (population) [BUI 05]. L'objectif d'automatisation se traduit par l'intégration des techniques d'extraction d'information (projeter sur les textes des automates repérant les variantes linguistiques exprimant concepts et relations) et des techniques d'apprentissage automatique (des textes annotés manuellement servent alors de corpus d'apprentissage). Le but est de trouver des régularités pour en dégager des classes sémantiques ou des patrons de fouille repérant ces classes et leurs relations. La prise en compte du web comme source de corpus ajoute évidemment de nouveaux défis, comme le volume des données à gérer, leur hétérogénéïté ou encore leur fiabilité.

## 1.3. Utilisation des ontologies en recherche d'information : état de l'art

### *1.3.1 Evolution des enjeux de la RI*

Dans un premier temps, les études en RI se sont focalisées sur les performances des moteurs de recherche, des bases de données servant au stockage de gros volumes de documents, des systèmes de classement documentaire et d'indexation. Arrivés à une certaine maturité, ces travaux ont eu l'ambition justifiée de mieux adapter les scénarios d'usage aux pratiques et aux besoins des utilisateurs [BOU OOa].

Depuis une dizaine d'années, les travaux menés en RI évoluent et sont largement influencés par l'intérêt porté à ce domaine par les communautés de l'IA et de l'IC. Il s'agit de ne plus traiter les textes comme des ensembles de mots astucieusement analysés, dont on étudie les régularités et les fonctionnements d'un point de vue statistique et quantitatif. L'enjeu est alors de doter les applications informatiques de la capacité de déterminer de quoi parle un document, de retrouver les documents ou parties de documents traitant d'un sujet particulier, de juger de la nouveauté d'une information, de répondre à des questions précises ou encore de constituer des dossiers thématiques, voire des modèles de connaissances [AUS 04b]. Cette tâche, maîtrisée par les documentalistes, suppose un outillage complexe, même pour l'humain en charge de l'élaborer et l'utiliser : définir des thésaurus fournissant les termes décrivant les centres d'intérêt d'un domaine, des langages documentaires pour annoter des documents en fonction de leur contenu, fixer un jeu de méta-données standards pour caractériser les documents et faciliter leur échange entre centres d'information, garder trace des préférences et centres d'intérêt des utilisateurs, etc.

Au cœur du projet du Web Sémantique [CHA 05], les ontologies sont considérées comme le moyen de disposer de modèles de connaissances partageables, consensuels et permettant de raisonner sur les connaissances. Les ontologies sont vues, en recherche d'information, comme une version élaborée et formelle des

thésaurus et langages documentaires, permettant d'élargir les possibilités de caractérisation des documents et des besoins en information des utilisateurs. Elles sont le creuset de mots-clés servant à définir des méta-données, à caractériser des documents ou les indexer.

### *1.3.2 Apports attendus des ontologies dans la recherche d'information*

La recherche d'information se tourne vers les ontologies et les modèles assimilés pour caractériser le contenu (la sémantique) des documents (ou de granules documentaires) et les besoins des utilisateurs et ce afin d'améliorer les approches classiques. L'exemple des applications de veille documentaire illustre la diversité des attentes vis à vis de l'utilisation des ontologies. On espère ainsi un meilleur ciblage des besoins (une expression claire de la question qui intéresse une communauté d'utilisateurs et la description des sources à analyser), une sélection plus pertinente des documents au sein d'une source donnée, un traitement plus efficace des documents (indexation et classification) et enfin la restitution de résultats vers des utilisateurs mieux ciblés par leurs centres d'intérêt [CAO 05].

L'objectif est de définir de nouvelles représentations des textes qui soient plus riches, plus précises et plus efficaces. Les approches classiques essaient de résoudre ces questions de manière endogène, en exploitant des statistiques relatives à l'usage des mots et leurs cooccurrences. L'utilisation de ces ressources pour une indexation sémantique revient à s'appuyer sur des relations que les concepts entretiennent dans un modèle pour donner du sens à l'information exprimée dans ces documents. Ainsi, de manière complémentaire au contenu des textes, on fait appel à des informations absentes des textes (mais explicites dans la ressource sous forme de « connaissances ») qui aideront à mieux en caractériser le contenu. Cette ressource doit donc offrir un vaste inventaire de termes, et les associer à des concepts, ainsi qu'un réseau sémantique riche reliant les concepts en fonction de points de vue sur ce qu'ils signifient. Différents types de traitements ont été imaginés pour exploiter ces ressources et qualifier l'information présente dans les documents.

Les questions à traiter se posent depuis le début de l'automatisation de la recherche d'information. Il s'agit d'abord de difficultés linguistiques, comme les phénomènes de polysémie et de variation lexicale. On espère améliorer l'indexation en distinguant les occurrences d'un terme qui correspondent à des sens différents, et en regroupant les mots synonymes ou les différentes formulations d'une même idée. D'autres limites sont liées à la formulation écrite des requêtes, souvent courtes. On cherche alors à les étendre ou à les reformuler automatiquement pour retrouver des documents n'utilisant pas exactement les termes de l'utilisateur mais répondant à sa recherche d'information. Un autre axe est de localiser l'information pertinente précisément au sein de documents structurés. Enfin, l'ontologie est envisagée aussi

pour mieux expliciter les centres d'intérêt des utilisateurs ou encore les paramètres d'utilisation des systèmes (support matériel, contexte, type d'utilisateur, etc.).

### *1.3.3 Ontologies pour l'indexation conceptuelle ou sémantique*

Certains chercheurs [MIH 00] différencient *indexation conceptuelle* lorsqu'on fait appel à des hiérarchies de concepts ou à des ontologies spécialisées de *l'indexation sémantique* lorsque l'indexation s'appuie sur une ressource lexicale ou sémantique. Cependant, nous considèrerons les deux termes comme équivalents par la suite. Rappelons les deux courants à l'origine de la diversité actuelle de ce que couvre l'indexation sémantique [HER 05a]. D'un côté, dans les travaux issus de la RI, les concepts et instances de concepts de l'ontologie sont choisis comme langage de représentation des documents ; il s'agit alors d'identifier les concepts indexant les granules, puis de pondérer ces concepts pour améliorer la discrimination[3]. D'un autre côté, dans les travaux issus du courant « Web Sémantique », les documents sont caractérisés à l'aide de méta-données, qui s'ajoutent au texte du document pour permettre de mieux y accéder ; on parle d'ailleurs plus *d'annotation* que *d'indexation*. Les ontologies servent alors de sources de méta-données structurées et formalisées pour décrire le contenu des documents, alors que d'autres méta-données identifient la localisation, le format ou la production du document.

Lorsque la RI s'intéresse à l'indexation sémantique pour des moteurs de recherche généraux, les ressources utilisées doivent couvrir la langue générale pour améliorer les performances (rappel et précision) des moteurs. Plusieurs travaux ont montré les limites de cette hypothèse : l'utilisation non contrôlée de ressources comme la base de données lexicales WordNet peut au contraire dégrader les résultats car la multiplication des concepts génère du bruit [BAZ 05a]. Des approches plus récentes améliorent les résultats grâce à des choix précis [BAZ 05b] : la représentation du document est un réseau sémantique composé de concepts de WordNet reconnus dans les textes à partir des termes les plus précis possible, étendu à des concepts reliés par des relations sémantiques. À l'inverse, des tâches comme la classification de veille ou l'aide à l'activité de veille supposent que l'utilisateur soit un spécialiste du domaine sachant formuler précisément ses centres d'intérêt. Le domaine à couvrir est généralement bien ciblé, ainsi que la nature ou les caractéristiques des documents recherchés.

Dans le cadre du Web Sémantique, l'indexation de textes est souvent reformulée comme le problème du repérage *d'instances de concepts* d'ontologies dans les textes alors que l'association de méta-données relève d'une caractérisation de l'information à un niveau différent. Pour repérer des instances de concepts dans des

[3] Pour un inventaire de ces techniques, consulter [HER 05], [BAZ 05] ou [NJO 05]

textes, l'analyse du langage naturel et les techniques d'extraction d'information offrent des perspectives prometteuses [BUI 05] : si on arrive à caractériser les contextes (lexicaux ou grammaticaux) d'apparition d'un concept dans un texte, soit manuellement soit par apprentissage à partir d'un échantillon de textes étiquetés à la main, on peut localiser dans des documents des phrases où se trouvent des concepts ou des instances de concepts. De ce fait, les processus d'indexation sont les symétriques de l'analyse de textes qui sert à construire les ontologies [AUS 05b]. Les mêmes recherches permettent donc de progresser sur ces deux fronts, et constituent une des clés actuelles de la réussite du Web Sémantique. Le coût de la construction de l'ontologie et de l'indexation est encore élevé. De ce fait, on trouve deux types de travaux : de grands projets à vocation générale pour construire des ressources universelles (comme WordNet) ou génériques (comme DOLCE [MAS 03]) ; des applications dans des domaines spécialisés et visant la recherche d'informations importantes, coûteuses et précises dans les textes.

### *1.3.4 Ontologie pour l'accès aux documents*

Dans le cycle de la RI, le pendant de la représentation des documents est la formulation de requêtes ou de besoins par les utilisateurs. Reste ensuite à apparier ces deux représentations : la requête et les documents. L'utilisation des ontologies renouvelle chacune de ces tâches, et leur utilisation laisse envisager de nouvelles manières de présenter les résultats de recherches ou les collections à explorer. Nous reprenons ici le panorama dressé dans [HER 05a].

L'approche classique suppose une interrogation formulée en langage libre. À partir du moment où une ontologie sert de langage pivot, plusieurs solutions ont été envisagées : proposer de formuler directement la requête à l'aide d'une formule logique dans le langage formel de l'ontologie (cf. PICSEL [ROU 03]), interroger le système en sélectionnant une combinaison de concepts, via une interface spécifique, ou encore autoriser une formulation en langage naturel traduite ensuite sous forme de concepts à partir des termes de l'ontologie. La requête exprimée à l'aide de concepts peut être enrichie de concepts proches au sens de l'ontologie [BAZ 05a] [NJO 05]. Elle peut être considérée comme une formule logique (conjonction de concepts) [ROU 03], un sous-réseau conceptuel [BAZ 05b], un graphe dont les nœuds sont des concepts et les arcs les relations existant dans l'ontologie entre ces concepts [GUA 99], un vecteur de concepts ou encore une simple liste de concepts.

Ensuite, l'appariement se fait entre la requête et les documents en évaluant la proximité entre la représentation de la requête et celle correspondant à chaque document. Ce calcul de similarité dépend de la représentation choisie. Il intègre le plus souvent une mesure de distance sémantique entre concepts [KHE 05]. Ce type de mesure s'appuie sur le nombre et la nature des relations entre concepts dans

l'ontologie (nombre d'arcs dans les graphes représentant document et requête dans OntoSeek [GUA 99]). Il s'agit là d'un domaine de recherche très actif car le calcul de distance sert également à comparer des ontologies entre elles, à les fusionner pour les réutiliser, ou encore à juger de leur pertinence par rapport à une collection de documents donnés.

Enfin, l'utilisation d'ontologies permet d'innover en définissant des interfaces originales de navigation dans les résultats. L'exemple de la catégorisation de documents est tout à fait illustratif. Pour regrouper les documents d'une collection selon des classes prédéfinies, les ontologies servent à faciliter l'expression des classes en fonction des préférences et des centres d'intérêt des utilisateurs. Dans le cas simple où les catégories sont définies par des termes organisés en hiérarchie, consulter la hiérarchie permet de balayer la collection selon ces catégories plus ou moins spécifiques. Plusieurs hiérarchies peuvent être croisées pour exprimer des points de vue plus élaborés [AUS 04a], ou comporter des concepts (au sens minimal de classes de termes synonymes).

Plus sophistiquée, la structure de thèmes de catégorisation peut être une ontologie. L'ontologie organise soit les documents, soit les méta-données. Dans [STU 04], un document indexé définit un concept à l'aide de propriétés et des termes associés. Le système calcule des rapprochements de documents à partir des recouvrements des termes les indexant. Dans [MAE 02b], l'ontologie classe les méta-données associées aux documents, et le rapprochement des documents se calcule sur la base de ces méta-données. Ces approches utilisent peu la pondération des concepts. L'intérêt des ontologies est ici de mettre à plat une représentation structurée couvrant un domaine particulier, qui permette de jouer sur les relations entre concepts, et d'en identifier la terminologie, pour couvrir le vocabulaire utilisé dans les documents et dans les besoins en information.

## 1.4. Recherche d'information à l'aide d'ontologies : retours d'expérience

Les expériences rapportées ici, auxquelles nous avons participé, traduisent la diversité des approches et des apports possibles des ontologies à la RI.

### *1.4.1 DocCore : Ontologies et recherche d'informations générales*

Afin d'évaluer l'apport de modèles conceptuels à la recherche d'information en amont d'un moteur de recherche général, deux études ont été menées successivement. L'une porte sur la reformulation de requêtes en exploitant les relations entre concepts [BAZ 03], l'autre sur la représentation de documents sous forme d'un réseau de concepts puis d'un arbre de concepts [BAZ, 05a].

**Reformulation de requêtes :** Une première expérience a donc consisté à utiliser une base de données lexicale pour la reformulation des requêtes utilisateurs. Pour assurer un gain dans les résultats retournés, un processus d'"expansion prudente" a été défini en amont d'un moteur de recherche. Ce processus, transparent à l'utilisateur, exploite d'abord la notion de concepts multi-termes pour désambiguïser les mots de la requête (au sens de WordNet). Il s'appuie ensuite sur les relations sémantiques entre concepts pour élargir la requête. Différents tests ont été effectués pour évaluer ce processus qui conduit à une amélioration significative de la pertinence des réponses fournies par le moteur. Les expérimentations ont été réalisées en utilisant le moteur Mercure développé à l'IRIT, WordNet comme base de données lexicales et Clef2001 comme collection de test [BAZ 03]. WordNet comporte un réseau de nœuds lexicaux, appelés Synsets, reliés par des relations rendant compte des liens entre termes dans la langue (liens d'hyperonymie, de méronymie, etc.).

Le travail précis mené au cours du DEA de M. Baziz montre que la nature des relations entre les nœuds du modèle conceptuel a une influence significative sur l'expansion de requête [BAZ 03]. Les requêtes étendues avec des concepts reliés à ceux de la requête initiale permettent d'améliorer les résultats par rapport à la requête d'origine sous certaines conditions. Le succès suppose de minimiser le nombre de concepts utilisés pour représenter une requête, et de n'exploiter que les relations est-un pour l'expansion (les relations de méronymie ou d'antinomie au contraire n'améliorent pas les résultats). Les modules d'expansion de requête en choisissant le type de relation ont été intégrés à la plate-forme de recherche d'information RFIEC[4], afin de pouvoir reproduire l'expérience sur d'autres corpus.

**Représentation sémantique de documents :** De manière symétrique à l'enrichissement de la requête, dans sa thèse, M. Baziz a défini une représentation sémantique des documents [BAZ 05b]. Cette représentation, appelée *noyau sémantique du document*, prend la forme d'un réseau de concepts jugés représentatifs du document. Pour identifier automatiquement ces concepts à partir d'une ressource générale ou « ontologie » (ici WordNet), l'approche consiste à projeter les documents sur cette ressource. Pour chaque document, les concepts le *représentant* sont choisis à partir des termes du texte. La proximité sémantique dans l'ontologie permet de désambiguïser les termes pour ne retenir qu'un concept parmi plusieurs candidats, en fonction de ses termes voisins en corpus et des différents mots de sa définition dans WordNet. Ensuite, les concepts sont pondérés (deux poids ont été étudiés : cf.idf inspiré du tf.idf, et le C_score, tenant compte des concepts reliés au concept étudié dans l'ontologie). Seuls les concepts d'un poids supérieur à un seuil sont retenus.

[4] http://www.irit.fr/RFIEC

Le *noyau sémantique du document* représente le contenu informationnel du document à l'aide de nœuds (les concepts désambiguïsés) et d'arcs (liens de similarité sémantique calculés à partir de relations présentes dans WordNet et d'une distance sémantique choisie). Le calcul de ce noyau, long et coûteux, est fait une fois pour toutes pour une distance donnée. Ainsi, la collection interrogée est représentée par l'ensemble des noyaux sémantiques des documents qui la composent. Lors de la recherche d'information, la requête est traduite sous forme de concepts et étendue selon les principes d'expansion prudente. Puis elle est comparée aux différents noyaux sémantiques pour identifier les documents les plus pertinents.

Six distances ont été comparées sur un jeu de test pour mesurer la proximité sémantique entre concepts. Il en ressort que la mesure de Resnik est la plus efficace sur la collection utilisée, combinée au calcul du C_Score [BAZ 05b]. L'ajustement des poids associés aux concepts s'avère d'un impact presque aussi important sur la qualité des résultats que le choix des concepts eux-mêmes. En effet, pour le moteur de RI utilisé, la représentativité des concepts (explicitée par leur pondération) est importante pour classer et comparer des documents répondant à une requête.

**Bilan :** L'approche proposée est originale dans la mesure où elle combine à la fois la richesse d'une représentation à base de concepts (le noyau sémantique) et l'efficacité de la pondération (à l'aide de C_Score) utilisée en RI pour rendre compte de l'importance des concepts dans les documents. Finalement, WordNet se prête bien aux évaluations menées sur des corpus de référence (en anglais) comme ceux des programmes CLEF et TREC, en particulier grâce à la richesse de son réseau de Synset, à la diversité des relations présentes et au vocabulaire associé à chacun des Synsets. Le fait de constituer les noyaux sémantiques de manière automatique est à la fois un atout (efficacité, transparence pour l'utilisateur) et une limite (difficulté de vérifier leur représentativité des documents). Une représentation graphique de ce noyau montre qu'il contient en général des concepts clés des documents. Cette étude a permis de reformuler la question de « l'apport des ontologies à la recherche d'information (en général) » en « dans quelles conditions les ontologies peuvent améliorer la recherche d'information ? ». Ces conditions concernent la tâche de RI étudiée, les ontologies adaptées (contenu, degré de formalisation, couverture du domaine) ainsi que des heuristiques pour exploiter les relations entre concepts.

### *1.4.2 Ontologies pour l'exploration documentaire et la veille scientifique*

Afin d'illustrer l'utilisation d'ontologies pour l'exploration de collections dans le cadre de la veille scientifique. N. Hernandez a défini un environnement d'exploration de collections (OntoExplo) à partir de hiérarchies de concepts de ce domaine et de concepts décrivant la tâche de veille [HER 05]. La place et le rôle de cette « ontologie » y sont envisagés sous un angle assez différent. Dans OntoExplo,

l'organisation hiérarchique des concepts selon un ou plusieurs points de vue joue un rôle privilégié. Par exemple, dans le domaine de l'astronomie, des articles scientifiques peuvent être rassemblés en fonction de critères comme les objets astronomiques dont ils parlent, des instruments de mesure mentionnés, des stations observatoires, des journaux dans lesquels ils sont parus, de leur date ou de leurs auteurs. Chaque critère définit un point de vue qui organise un ensemble de concepts plus précis. Le choix de plusieurs critères permet de constituer des groupes de documents traitant de sujets plus ou moins précis, d'affiner les classes de documents en fonction des concepts caractérisant leur contenu.. Ensuite, un environnement de visualisation, OntoExplo, présente plusieurs hiérarchies pour faciliter la focalisation sur des documents particuliers et en assurer la consultation rapide. Les hiérarchies de concepts sont vues comme un guide pour naviguer dans l'espace d'information que constitue la collection de documents. L'utilisateur choisit des concepts, la collection est réorganisée en fonction des points de vue associés, et l'utilisateur peut alors explorer la collection et naviguer entre les documents.

Le modèle de données de la structure ontologique a été défini de manière ad hoc pour réduire les phénomènes classiques qui dégradent les résultats : ambiguïté de mots polysémiques, différence de vocabulaire entre les textes et les formulations des utilisateurs ou encore mauvaise gestion des variations de forme, des ellipses, etc. Pour élaborer les hiérarchies de concepts, les experts du domaine sélectionnent les points de vue, choisissent les concepts qui les intéressent et les organisent en hiérarchies. Ensuite, c'est l'analyse de corpus qui assure un enrichissement terminologique de ces hiérarchies. Les concepts et le vocabulaire associé servent aussi de langage pour exprimer le besoin en information.

L'indexation à l'aide des concepts des hiérarchies est assez immédiate. Elle exploite les termes associés aux concepts et des traitements linguistiques élémentaires comme la lemmatisation. Il paraît difficile d'imaginer qu'une ontologie unique organise et associe entre elles ces hiérarchies. En effet, la cohérence de ce modèle requiert de prendre un parti permettant d'unifier les différentes dimensions, tout en respectant des principes de structuration ontologique. Une ontologie constitue donc une représentation du domaine trop complexe pour être présentée directement lors de l'exploration de collections. Chaque hiérarchie correspond à une dimension d'analyse, à un point de vue sur le domaine. Elle est représentée par un arbre de concepts reliés selon une relation unique (est-un ou une autre relation). L'organisation hiérarchique facilite une visualisation graphique et la navigation en focalisant la recherche sur des ensembles plus ou moins restreints de documents. La présentation de l'ensemble des concepts des hiérarchies ou de l'ontologie donne à l'utilisateur une idée du contenu des documents de la collection.

La figure 4 présente un exemple d'exploration de corpus d'atricles pour une tâche de veille en astronomie. L'interface permet de visualiser, pour un article

donné, l'ensemble des méta-données intéressant l'utilisateur : noms des auteurs, date et revue de publication (partie de gauche de l'écran) mais aussi l'ensemble des concepts du domaine abordé dans l'article : système solaire, comète (partie de droit de l'écran).

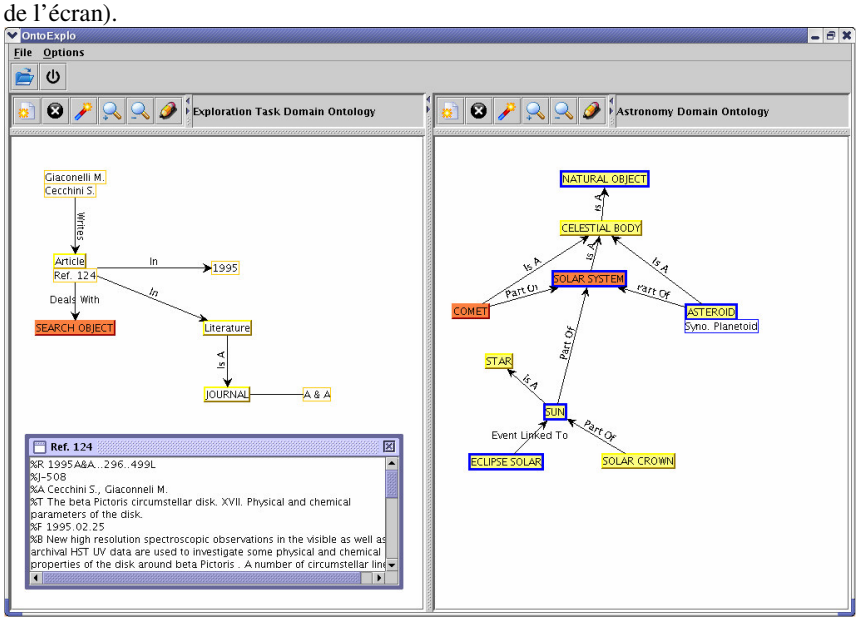


**Figure 4** : *Visualisation de la connaissance établie pour un article à partir du corpus*

Ce genre d'approche suppose de s'intéresser à des domaines stables, dans lesquels les connaissances évoluent peu. Ou alors il faut définir un protocole de maintenance de l'ontologie et de définition de nouveaux points de vue au fur et à mesure que les connaissances du domaine évoluent ou que la collection s'enrichit de nouveaux documents.

### *1.4.3 Ontologies pour l'annotation sémantique*

Le projet Arkeotek[5] vise une meilleure gestion de l'archivage de monographies, d'articles multilingues et de données (dessins, photographies, …) liés à l'archéologie des techniques [GAR 04]. Afin de constituer des bases de connaissances, les documents sont structurés selon les principes du logicisme et les travaux de Gardin, en identifiant les différentes briques élémentaires de raisonnement, sous forme de propositions, et la contribution des unes aux autres pour produire des inférences. Les

[5] www.arkeotek.org

documents sont ensuite transformés en documents électroniques par le biais du format SCD [ROU 04]. Ce format se caractérise par l'édition structurée des textes scientifiques en fragments hiérarchisés (propositions interprétatives et antécédentes). Il se veut avant tout une grille d'organisation des écrits afin d'en consolider la rigueur argumentaire et ainsi la qualité scientifique. Le format peut se traduire par une DTD XML où les balises précisent le rôle des paragraphes dans l'expression d'une argumentation scientifique. Cette fragmentation permet, d'une part, un accès plus facile et rapide aux constructions scientifiques et, d'autre part, une indexation raisonnée des bases de données documentaires en fonction des raisonnements qu'elles présentent, répondant par là aux questions d'archivage de masses de données scientifiques. À terme, les bases ainsi constituées favorisent le cumul des connaissances au sein des SHS, à condition toutefois de prévoir des outils pour les interroger [AUS 05a].

Pour répondre à ce besoin, des outils de requête permettant d'accéder rapidement aux résultats disséminés au sein de nombreuses publications ont été définis. Le choix retenu consiste en l'indexation sémantique des publications à l'aide d'une ontologie du domaine, elle-même construite à partir de ces documents. L'indexation s'appuie alors sur une double caractérisation du contenu des documents : l'une porte sur l'argumentation scientifique à laquelle correspond chaque fragment de texte, et correspond à la structuration selon le format d'édition scientifique SCD ; l'autre vient enrichir les documents structurés par une représentation explicite des connaissances du domaine sur lesquelles porte chaque fragment de document, sous la forme de concepts. Ces concepts, tirés d'une ontologie du domaine concerné, constituent un index sémantique des documents.

L'ontologie est construite à partir d'une analyse linguistique du contenu des documents à indexer. Ainsi, en retour, on utilise les liens identifiés entre termes et textes pour établir des liens entre les concepts de l'ontologie et les textes. Cette association entre concepts et fragments de textes est supervisée. Elle tient compte du typage SCD des paragraphes. La disponibilité de ces liens facilite l'indexation.

Un prototype du système d'interrogation de la base de documents est en cours de développement. Il comprend deux parties. *L'environnement auteur* est destiné à construire la représentation enrichie des textes de la base. Un module guide la construction et la maintenance de l'ontologie à partir de documents structurés, un autre gère l'indexation supervisée de nouveaux documents à l'aide de l'ontologie. *L'environnement utilisateur* permet d'interroger la base documentaire ainsi indexée. L'utilisateur peut formuler une requête, que le système projette sur la base documentaire et pour laquelle il présente des paragraphes pertinents.

*Nature de l'ontologie* : L'application visée détermine fortement le niveau de détail de l'ontologie. À partir du moment où l'on estime que les utilisateurs ne feront

pas de différence entre deux termes au moment d'interroger la base documentaire, un seul concept est présent dans l'ontologie, auquel tous les termes plus précis ou proches sont associés. De ce fait, l'ontologie comporte une composante terminologique riche, mais peu de concepts détaillés. Cette ontologie n'est pas formalisée en logique.

*Mode d'indexation* : La solution envisagée privilégie une analyse automatique des textes et un travail de modélisation poussé au moment de construire l'ontologie. Dans ce modèle, les liens des termes vers leurs occurrences sont conservés. Ils sont exploités lors de l'indexation. Le processus d'indexation est supervisé : le système propose un ensemble de concepts à associer à chaque paragraphe en fonction des termes présents et d'heuristiques liées au format SCD ; l'utilisateur vient ensuite les modifier ou les valider. Le système implémente des solutions pragmatiques aux questions soulevées qui n'ont pas encore été évaluées. De plus, la nature de la pondération associée aux concepts ainsi que le processus d'appariement restent à définir. Cependant, cette étude mérite d'être mentionnée ici car elle illustre bien la nécessité de s'adapter aux contraintes de la tâche de RI ainsi que l'atout de la prise en compte de la structure des documents.

## 1.5. Bilan et conclusion

### *1.5.1 Adéquation des ressources sémantiques à la recherche d'information*

A travers ces états de l'art et études de cas, nous voulons souligner que l'intérêt d'utiliser des ressources sémantiques pour la recherche d'information n'est ni immédiat ni systématique. Il requiert une réflexion approfondie sur les exigences de la tâche visée, de manière à définir le type de RTO à utiliser (sa complexité, sa généralité, son degré de formalisation, etc.) et les modalités de son utilisation. A chaque étape du cycle de recherche d'information, de nombreuses alternatives sont possibles et doivent être évaluées avant de retenir une approche particulière.

Nous insistons sur quelques leçons qui peuvent être tirées des études réalisées à ce jour. En premier lieu, concernant le contenu de la RTO, sa richesse terminologique est déterminante. La représentation de textes ou de requêtes sous forme de concepts impose d'inventorier le plus grand nombre de variantes de formes des concepts, de synonymes, paraphrases ou termes associés. De manière complémentaire, un autre résultat concerne la nature des relations exploitées lors de la reformulation de requête ou l'élargissement de la représentation des textes. Ces relations doivent être utilisées avec prudence. Seule des relations à la sémantique maîtrisée et précise, « sûres » comme est-un, produisent un gain, à condition de ne les exploiter que sur un seul niveau autour de chaque concept.

Il se dégage de nos analyses que les représentations des documents construites à l'aide de concepts sont à inventer ou adapter pour chaque type de tâche de recherche d'information. Elles peuvent tirer profit tirer profit de la richesse terminologique et de la richesse des relations de la RTO utilisée. La diversité des problèmes de RI justifie la diversité des représentations à proposer (graphes, réseaux, vecteurs de termes ou de concepts, etc.). De même, pour construire ces représentations, c'est-à-dire pour indexer ou annoter les documents, il semble indispensable de combiner astucieusement les approches statistiques et sémantiques. Enfin, pour définir l'appariement entre requête et document, le choix du mode de calcul de proximité et de la distance sémantique dépend étroitement de la représentation retenue.

### *1.5.2 Perspectives de recherche*

Parmi les perspectives de recherche qui se dégagent, les sujets d'actualité concernent l'automatisation du processus de construction de la RTO et du processus d'indexation sémantique par une fouille de textes qui combine des algorithmes de TAL, d'extraction d'information et d'apprentissage. Il est classique de rappeler que ces approches gagnent à combiner des mesures statistiques et des approches linguistiques. Avec la généralisation de l'utilisation de XML, il est prometteur d'exploiter aussi la structuration des documents, et de constituer des patrons de recherche d'information ou de chercher des corrélations tenant compte de l'information présente et de sa mise en forme ou de sa place dans la structure du document. Enfin, l'utilisation des ontologies et des RTO en recherche d'information vient alimenter le débat sur ce que doivent être les ontologies, sur la faisabilité de définir des ressources stables et partageables, mises à disposition sur le web. Le succès de leur utilisation dépend d'une bonne couverture des documents par la ressource, ce qui suppose qu'elle puisse être soit très générale, soit ajustée et mise à jour régulièrement. Elle souligne donc l'importance de la maintenance de la composante terminologique et la nécessité de se doter de moyens d'identifier de nouvelles instances en corpus. Le statut de cette ressource n'est donc pas simplifié ou clarifié par l'informatisation, on retrouve là des problèmes bien connus des sciences de l'information dans la gestion des langages documentaires.

## 1.6. Bibliographie

[AHM 95] AHMAD K., HOLMES-HIGGIN P.R., "SystemQuirk : a unified approach to text and terminology", in *Terminology in advanced Microcomputer Applications*. Proc. of the 3rd TermNet Symposium : recent advances and user reports, Vienna, Austria, 181-194, 1995.

[ARP 03] ARPIREZ J., CORCHO O., FERNANDEZ-LOPEZ M., GÓMEZ-PÉREZ A., "WebODE in a nutshell", in AI Magazine, 24(3), pp 37-48, 2003.

[AUS 01] AUSSENAC-GILLES N., CONDAMINES A., « Entre textes et ontologies formelles : les bases de connaissances terminologiques ». *Ingénierie et capitalisation des connaissances.* Eds. M. Zacklad, M. Grundstein. Paris : Hermès. Traité IC2. 153-177. 2001

[AUS 03] AUSSENAC-GILLES N., CONDAMINES A., *Action spécifique STIC « Corpus et Terminologie » ASSTICCOT (AS 34). Rapport final.* Rapport IRIT/2003-23-R. 2003.

[AUS 04a] AUSSENAC-Gilles N., MOTHE J., "Ontologies as Background Knowledge to Explore Document Collections". *RIAO 2004*, 129-142. 2004.

[AUS 04b] AUSSENAC-GILLES N., CONDAMINES A. « Documents électroniques et constitution de ressources terminologiques ou ontologiques ». *Revue Information, Interaction, Intelligence I3*. Numéro spécial sur le document numérique. Eds CHARLET J. et SALAÜN J.-M. 4(1):75-94. 2004.

[AUS 04C] AUSSENAC-GILLES N., BIEBOW B., SZULMAN S., « Modélisation du domaine par une méthode fondée sur l'analyse de corpus ». *In Ingénierie des Connaissances*. R. Teulier, P. Tchounikine et J. Charlet Eds. Paris : L'harmattan. 2004.

[AUS 05a] AUSSENAC-GILLES N., ROUX V, de SAIZIEU B., BLASCO P., Ontologies dédiées à la consultation de documents structures selon un modèle logico-sémantique. In *Actes du colloque de clôture du programme Société de l'Information*. Lyon (F), mai 2005.

[AUS 05b] AUSSENAC-GILLES N., « Méthodes ascendantes pour l'ingénierie des connaissances », Mémoire d'habilitation à diriger des recherches de l'université Paul Sabatier (Toulouse 3). Déc. 2005

[BAC 04] BACHIMONT B. Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle. Mémoire d'habilitation à diriger des recherches de l'Université Technologique de Compiègne. Janvier 2004.

[BAZ 03] BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., « Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques ». *Revue des Sciences et Technologies de l'Information (RSTI)* série ISI, Hermes : Paris, V. 8, N. 4/2003, 113-136, Déc. 2003.

[BAZ 05a] BAZIZ M., Indexation conceptuelle guidée par ontologie pour la recherche d'informations. Thèse de doctorat de l'Université Paul Sabatier, Toulouse. Déc. 2005.

[BAZ 05b] BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., "A Conceptual Indexing Approach based on Document Content Representation". *In proceedings of COLIS 2005 Context: nature, impact and role.* Univ. Of Strathclyde, Glasgow (UK), July 2005. F. Crestani and I. Ruthven (Eds.): LNCS 3507. Berlin : Springer-Verlag, 2005. 171-186.

[BER 99] BERNERS LEE T., HENDLER J., LASSILA O., « Semantic Web ». *Scientific American*. 1999.

[BOU 00a] BOUGHANEM M., « Contribution à la formalisation et à la spécification des systèmes de Recherche et de Filtrage d'Information ». Habilitation à diriger des recherches. Université Paul Sabatier, Toulouse. Nov. 2000.

[BOU 00b] BOURIGAULT, D. & JACQUEMIN, C., Construction de ressources terminologiques, in J.-M. Pierrel (éd), *Ingénierie des langues*, Traité I2C, Paris, Hermes. 2000.

[BOU 02] BOURIGAULT D. , UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, 75-84, 2002.

[BOU 04] BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA*). Numéro spécial sur les Techniques Informatiques et Structuration de Terminologies. PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. **18** (1) : 87–110. 2004

[BUI 04] BUITELAAR P., OLEJNIK D., SINTEK M., A Protégé Plug-In for Ontology Extraction from Text based on Linguistic Analysis. In *Proceedings of the 1st European Semantic Web Symposium*. Héraklion (Greece), May 2004.

[BUI 05] BUITELAAR P., CIMIANO P. and MAGNINI B., *Ontology Learning from Text: Methods, Evaluation and Applications*. Volume 123, Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

[CAO 05] CAO T-D. DIENG-KUNTZ R., FIES B., BOURDEAU M., « Vers un système d'aide à la veille technologique guidé par une ontologie ». *Actes des Posters de la Conférence Ingénierie des Connaissances*, 2005 .

[CHA 02] CHARLET J., L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Pierre et Marie Curie. Déc. 2002.

[CHA 05] CHARLET J., LAUBLET P., REYNAUD C., (Eds.) Numéro spécial « Web sémantique », *Revue Information, Interaction, Intelligence I3*, Cépadues-Editions, 2005.

[CIM 05] CIMIANO P., VÖLKER J., "Text2Onto, a framework for Ontology Learning and data-driven Change Discovery", *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems,* Lecture Notes in Computer Science, 3513 : 227-238, 2005

[GAR 04] GARDIN, J-C., ROUX, V., "The Arkeotek project: a European network of knowledge bases in the archaeology of techniques". *Archeologia e Calcolatori*, **15**, 25-40. 2004.

[GOM 04] GÓMEZ-PÉREZ A., FERNÁNDEZ-LÓPEZ M, CORCHO O., *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Verlag, London. 2004.

[GRU 91] GRUBER T.R., "The role of common ontology in achieving sharable, reusable knowledge bases". *Proc. Of the 2nd Int. Conference on the Principles of Knowledge Representation and reasoning*. Morgan Kaufmann, San Mateo (Ca, USA). 601-602, 1991.

[GRU 93] GRUBER T. R, "Translation approach to portable ontology specifications". *Knowledge Acquisition,* **5**, 199-220. 1993.

[GUA 99] GUARINO N., MASOLO C., VETERE G., "OntoSeek: Content-Based Access to the Web". *IEEE Intelligent Systems*. 70-80. May-June1999.

[GUA 00] GUARINO N., WELTY C., "A formal Ontology of Properties". In Dieng R., Corby O. (Eds.) 12th *International Conference in Knowledge Engineering and Knowledge Management* (EKAW'00). Juan les Pins (F). LNAI 1937. Springer-Verlag. 97-112. 2000.

[HER 05a] HERNANDEZ N., Ontologies de domaine pour la modélisation du contexte en recherche d'information. Thèse de doctorat l'université Paul Sabatier, déc. 2005.

[HER 05b] HERNANDEZ N., J. MOTHE J., S. POULAIN S., "Accessing and mining scientific domains using ontologies: the OntoExplo System". *SIGIR*, 607-608, 2005.

[KAS 02] KASSEL G., « OntoSpec : une méthode de spécification semi-formelle d'ontologies ». *Actes des 13e journées francophones d'Ingénierie des Connaissances* (IC). 75-87. 2002.

[KHE 05] KHELIF K., Web sémantique et mémoire d'expériences pour l'analyse de transcriptome , Thèse en STIC de l'université de Nice-Sophia Antipolis. Mars 2006.

[LAS 01] LASSILA O., Mc GUINESS D. The role of frame-based representation on the Semantic Web, Technical report KSL-01-02, Knowledge Systems Laboratory, Stanford University, 2001.

[MAE 02a] MAEDCHE A., *Ontology learning for the Semantic Web*. Kluwer Academic Publisher. 2002.

[MAE 02b] MAEDCHE A., ZACHARIAS V., « Clustering Ontology-Based Metadata in the Semantic Web », *Principles of Data Mining and Knowledge Discovery (PKDD-2002),* Lecture Notes in Computer Science 2431, Springer, Berlin, 348-360, 2002.

[MAS 03] MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. and SCHNEIDER L., The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. WonderWeb Deliverable D18, Final Report (vr. 1.0, 31- 12-2003). 2003.

[MIH 00] MIHALCEA R. , MOLDOVAN D.I., "Semantic Indexing using WordNet Senses", *in Proc. Of ACL Workshop on IR & NLP*. 2000.

[MOR 99] MORIN E. , Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques, *Traitement Automatique des Langues*, **40** (1), 143-166. 1999.

[MOT 00] MOTHE J., Recherche et exploration d'informations - Découverte de connaissances pour l'accès à l'information, Habilitation à diriger des recherches. Université Paul Sabatier, Toulouse. Nov. 2000.

[NAU 06] NAUER E., RICHARD A., DERRIERE S., GENOVA F., NAPOLI A., TOUSSAINT Y., « Construction d'une ontologie de descripteurs UCD en astronomie », à paraître dans les *actes des 17e journées francophones d'Ingénierie des connaissances* 2006 .

[NJO 05] NJOMGUE SADO W., Indexation dees documents dans un référentiel métier avec approche ontologique : le système MAID au sein de l'intranet de Suez-Environnement. Thèse de doctorat de l'Université Technologique de Compiègne. 2005.

[ROU 03] ROUSSET M.-C., BIDAULT A., FROIDEVAUX C., GAGLIARDI H., GOASDOUE F., REYNAUD C., SAFAR B., « Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL ». *Revue I3 (Information-Interaction-Intelligence).* Cépaduès Editions, Toulouse. 2 (1). 2002.

[ROU 04] ROUX V., BLASCO P., « Faciliter la consultation de textes scientifiques. Nouvelles pratiques éditoriales ». *Hermès, Critique de la raison numérique*, CNRS éditions, 39, 151-159. 2004.

[STA 04] STAAB S., STUDER R. Eds., *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

[STU 98] STUDER R., BENJAMINS R., FENSEL D. Knowledge Engineering: principles and methods. *IEEE transactions on Data and Knowledge Engineering*. **25** (1-2): 161-197. 1998.

[STU 04] STUCKENSCHMIDT H., VAN HARMELEN F., DE WAARD A., SCERRI T., BHOGAL R., VAN BUEL J., CROWLESMITH I., FLUIT C., KAMPMAN A., BROEKSTRA J., VAN MULLIGEN E., "Exploring large document repositories with RDF technology: the DOPE project*",Intelligent system*, IEEE, **19** (3), 34- 40, 2004.

[SZU 02] SZULMAN, S., BIEBOW B., AUSSENAC-GILLES N., « Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE ». *Traitement Automatique de la Langue (TAL).* Numéro spécial « Structuration de Terminologie ». Eds A. Nazarenko, T. Hammon. **43** (1). Hermès : Paris : 103-128. 2002.

[SZU 04] SZULMAN S., BIEBOW B., « OWL et TERMINAE ». *Actes des 15° journées francophones d'Ingénierie des Connaissances (IC 2004)* Lyon (F.), Presses Universitaires de Grenoble : 41-52, 2004.

[USH 96] USCHOLD M. M., GRUNINGER M. "Ontologies: principles, methods and applications". *Knowledge Engineering Review*. **11** (2). 93-155. 1996.