Chapitre 1

Le web sémantique, quel renouvellement pour la recherche d'information ?

"*Humankind has not woven the web of life.*
We are but one thread within it.
Whatever we do to the web, we do to ourselves.
All things connect."
Chief Seattle, 1854

1.1. Introduction

Le Web Sémantique est un programme de recherche énoncé il y a près de 10 ans pour pousser plus loin l'exploitation du contenu du web par des applications informatiques et pour mieux répondre aux besoins des utilisateurs. Ce programme fait l'hypothèse de laisser le web en l'état et d'en gérer des versions enrichies d'une couche dite « sémantique », de manière normalisée et standardisée, au moment de l'exploiter. Il a débouché sur la proposition de technologies, de formats, de structures de données (en particulier les ontologies) à intégrer dans les applications du web. Or au cœur de ces applications, se trouve presque systématiquement une phase de recherche d'information (RI) qui doit être capable de gérer cette « sémantique ». La recherche d'information s'est donc renouvelée pour répondre à ce besoin, et évaluer le gain potentiel apporté par les technologies et ressources du web sémantiques.

Cet article rappelle les objectifs et champs d'investigation du web sémantique, puis il présente la notion d'ontologie, à la fois dans sa définition restrictive fournie par le W3C et dans sa diversité d'usage en ingénierie des connaissances et dans les systèmes d'information. A partir de là, nous développerons quelques spécificités de la recherche d'information dans le cadre du web sémantique, et en particulier la contribution des ontologies. Nous insisterons plus sur les processus d'annotation sémantique, de reformulation de requête et d'accès aux documents spécialisés. Pour finir, nous dresserons un bilan des résultats obtenus, les nombreuses perspectives qui se dessinent, en particulier la complémentarité avec les technologies mettant en valeur les usages, regroupées sous ce qui est appelé le Web 2.0.

1.2. Que recouvre le web sémantique aujourd'hui ?

1.2.1 Le projet du Web Sémantique

Voilà près de 10 ans que Tim Berners Lee a posé les bases du Web Sémantique. En 2001, il énonçait ainsi [BER 01a] :

"The Web of data (and connections) with meaning in the *sense that a computer program can learn enough about what the data means to process it. . . . Imagine what computers can understand when there is a vast tangle of interconnected terms and data that can automatically be followed."*

Le Web sémantique serait donc un vaste espace d'échanges de ressources *entre machines* permettant l'exploitation de grands volumes d'informations et de services variés, *aidant les utilisateurs en les libérant d'une (bonne) partie de leur travail de recherche, et de combinaison de ces ressources.*

Cette proposition visionnaire s'appuie sur le succès considérable d'Internet tel qu'il est, dont le potentiel fascine. Elle vise à profiter aussi de la numérisation croissante des collections et de la généralisation de la production de documents au format standardisé ou structurés à l'aide de XML. Cette proposition vient répondre aux limites perçues à l'époque, à savoir un utilisateur livré à lui-même, disposant de peu de moyens pour accéder rapidement à l'information qu'il recherche, pour repérer des informations précises, dont il doit prendre en charge lui-même l'interprétation ou l'utilisation dans son propre système d'information.

Une intuition sous-tend l'annonce de Tim Berners Lee : l'informatique doit être capable d'offrir plus de services à partir du web, des fonctionnalités plus efficaces et conviviales, faisant appel, si besoin, à des ressources présentes sur plusieurs sites et enchaînant des traitements sur ces ressources, pour les intégrer directement dans les

systèmes d'information locaux. Les objectifs ciblés dépassent ceux de la recherche d'information : il s'agit non seulement de localiser des informations, mais aussi de les manipuler, de les traiter en tant que données, informations ou même connaissances, pour des tâches ciblées. Par exemple, on ne se contentera pas de « lister tous les sites web vendant des VTT » mais bien de constituer un portail comparatif des produits disponibles chez différents fournisseurs, permettant de passer commande chez n'importe quel d'entre eux de manière transparente pour l'utilisateur, ou encore de faire une aide en ligne pour composer un VTT à partir de pièces commandées chez plusieurs fournisseurs.

Ce projet a pu être construit parce que plusieurs technologies, arrivées à maturité, semblaient des pistes prometteuses pour enrichir le potentiel du web : l'extraction d'information, le traitement automatique des langues, la gestion de méta-données et leur utilisation pour l'annotation, l'adaptabilité des interfaces aux matériels et aux usages, la capacité à envoyer des applications (des agents logiciels) et à les faire exécuter sur des serveurs, la gestion performante de machines mise en réseau et l'exécution répartie de services sur des « grilles », la recherche d'information dans sa diversité, mais aussi la capacité à gérer et modéliser des connaissances, la structuration des documents et la standardisation des formats des supports des différents médias (textes, images, vidéo, etc.). Un programme de recherche a donc été bâti sur la base de ces technologies, dont le but est énoncé en 2001 par le W3C [BER 01b] :

"*The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The mix of content on the web has been shifting from exclusively human-oriented content to more and more data content. The Semantic Web brings to the web the idea of having data defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications. For the web to reach its full potential, it must evolve into a Semantic Web, providing a universally accessible platform that allows data to be shared and processed by automated tools as well as by people*."[1]

[1] Le Web Sémantique est une extension du web actuel dans laquelle l'information prend un sens bien défini, permettant mieux aux ordinateurs et aux utilisateurs de coopérer. La combinaison des différents contenus sur le web est passée d'un contenu orienté vers les hommes à un contenu constitué de données (informatiques). Le Web Sémantique ajoute au web l'idée de disposer de données définies et mises en relation de manière à pouvoir les utiliser plus efficacement pour les retrouver, les automatiser, les intégrer et les réutiliser au sein de diverses applications. Pour que le web atteigne son plein potentiel, il doit évoluer vers un Web Sémantique, fournissant une plate-forme universellement accessible qui permette le partage des données et leur traitement par des logiciels aussi bien que par les humains.

1.2.2 La mis en oeuvre du web sémantique

Pratiquement, le W3C s'est focalisé sur deux aspects technologiques du projet du Web Sémantique, rappelés sur son site (http://www.w3.org/2001/sw/) : définir des formats d'intégration et de combinaison de données issues de diverses sources ; garantir que ces données fasse sens auprès des utilisateurs, et que les langues assurent un lien vers des connaissances qui aient le même « sens » pour le système et l'utilisateur, qu'elles « renvoient aux mêmes objets du monde ». Un panorama des travaux effectués dans la cadre du web sémantique a été dressé en 2003 par un groupe de chercheurs français à la demande du CNRS. Leur bilan [CHA 04a] confirme cette dualité : le web sémantique est à la fois *une infrastructure*, identifiée par un ensemble de technologies et par leur mise en réseau, mais aussi *le contenu de cette infrastructure*, à savoir des sites, des ressources, des données, et surtout *des applications* élaborées de recherche et de traitement de l'information.

En tant *qu'infrastructure*, le web sémantique doit permettre d'utiliser des connaissances formalisées (sans que l'on sache jusqu'à quel degré) et des étiquettes en plus du contenu informel et multi-média actuel du web. L'objectif est de dépasser les modes classiques de recherche d'information, de pouvoir combiner des informations provenant de plusieurs sites ou sources pour répondre à des besoins élaborés (comme composer un séjour touristique en gérant à la fois le séjour, les transports, les activités et visites, etc.), de s'adapter aux préférences des utilisateurs, de leur proposer des informations pertinentes plutôt que des les aider à les chercher, etc. Cette infrastructure doit permettre d'abord de localiser, d'identifier et de transformer des ressources du Web de manière robuste et valide, tout en étant accessible à une plus grande diversité d'utilisateurs.

Les clés du succès de ce projet sont triples [LAU 04] : tout d'abord, il faut un *effort technologique de standardisation* des formats d'échange, de codage, des représentations et des langages de traitement des données et des connaissances, tout cela pour dépasser l'hétérogénéité des formats, des données sur les sites et des contenus des ressources ; ensuite, il faut des données, *des ressources* (sémantiques et conceptuelles), des algorithmes capables d'associer cette dimension sémantique au web existant ; enfin, il faut entrer dans un cycle vertueux de *développement d'applications* et de ressources qui démontre que le coût (inévitable) ajouté par la dimension sémantique est compensé par des bénéfices significatifs pour tous, surtout et y compris pour ceux qui auront mis des ressources à disposition ou qui auront pris en charge des annotations.

1.2.3 Les réalisations technologiques du web sémantique

Aujourd'hui, l'effort technologique est fait et la volonté de disposer de ressources consensuelles est réelle : le W3C a mis en place plusieurs groupes de travail qui ont défini des standards compatibles avec XML. Ce choix se justifie par la volonté que les connaissances et informations ainsi représentés soient associés aux documents comme des méta-données, et traités de manière homogène aux marques décrivant la structure des données et décrite en XML (fig. 1). La communauté scientifique d'intelligence artificielle et d'ingénierie des connaissances dans son ensemble a adhéré à ce projet et l'a adopté avec enthousiasme. Preuve en est l'explosion du nombre de travaux, de conférences et de revues sur le Web Sémantique depuis 1999.

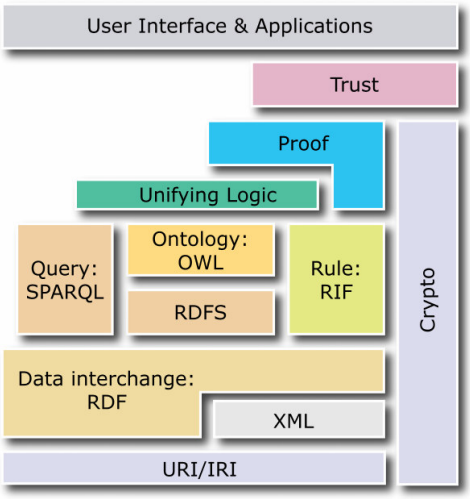

Figure 1 : *Les différentes couches du « gâteau » des technologies du web sémantique (http://www.w3.org/2001/sw/)*

Ainsi, ont été définis de nouveaux outils pour construire, adapter, comparer, fusionner, retrouver, vérifier la validité, utiliser et réutiliser des ressources sémantiques. La compatibilité avec XML conduit à identifier les données par des URI et à les définir au sein d'espaces de noms repérables sur le web [BER 98]. Pour la représentation des ontologies, une gamme de langages a été spécifiée [BAG 04]. Le format d'échange de données le plus simple, le Resource Description Framework (RDF), propose de représenter les données sous forme de triplets, généralisant ainsi la notion d'hyperlien entre ressources du web [BRI 99]. Pour mieux structurer et typer les données ainsi mises en relation, RDFs permet de définir des schémas RDF,

à savoir un vocabulaire réservé précisant la nature des ressources ou les noms des relations [BRI 02]. RDFs fournit des primitives de représentation telles que la notion de classe, de propriété et de relation hiérarchique entre classes. Ce langage convient pour représenter des ontologies simples. Enfin, le langage qui est devenu un standard de fait pour les ontologies est OWL, le « Web Ontology Language », qui enrichit RDFs de notions inspirées des frames et des logiques de description. OWL a été pensé de manière à disposer d'un langage plus ou moins riche de représentation des connaissances, et permettant de produire des raisonnements sur ces représentations [EUZ 04]. Pour cela, trois formalisations de OWL[2] sont disponibles (OWL-light, OWL-DL et OWL-full). Au-delà des langages, des standards portent également sur l'identification et la formalisation de types de méta-données par domaine, ou à l'adaptation de standard existant comme le Dublin Core pour les méta-données documentaires ou LOM pour le e-learning.

Ces langages ont été ensuite spécialisés pour être utilisés au sein d'agents logiciels, de services web et pour annoter des sites, des documents multi-média ou des données. Les web services offrent un champ de recherche à part entière, conduisant à définir plusieurs formats comme OWL-ws [KEL 04] WSDL (Web Service Description Language) et SAWSDL (Semantic Annotations for WSDL). Enfin, des logiciels ont été développés pour assurer, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ressources. Au niveau le plus haut, des mécanismes de protection (droits d'accès, d'utilisation et de reproduction) et d'évaluation du degré de fiabilité des connaissances sont requis.

Par effet de bord, on peut considérer que les ressources existent et qu'il y aurait matière à disposer d'un sous-ensemble intéressant du web au niveau sémantique. Ce qu'il manque le plus pour faire vivre et valider le projet du Web Sémantique, ce sont des applications « phare ». Bien sûr, le développement d'applications intégrant les standards du web sémantique, devrait être plus facile et à réduire le coût de services développés jusque là de manière souvent ad-hoc. Or l'utilisation et l'acceptation à l'échelle du web des technologies définies comme standards posent de nouveaux problèmes et défis aux chercheurs de l'intelligence artificielle (IA) et de l'ingénierie des connaissances (IC) : changement d'échelle dû au contexte de déploiement, le web et ses dérivés (intranet, extranet), nécessité d'un niveau élevé d'interopérabilité, ouverture, standardisation, diversités des usages, distribution bien sûr et aussi impossibilité d'assurer une cohérence globale [LAU 04]. Comme l'écrit, T. Berners-Lee, le web sémantique est ce que nous obtiendrons si nous réalisons le même processus de globalisation sur la représentation des connaissances que celui que le Web fit initialement sur l'hypertexte.

---

[2] http://www.w3.org/2001/sw/WebOnt/impls

1.2.4 La recherche d'information dans le cadre du Web Sémantique

Dans ce contexte nouveau, la recherche d'information (RI) continue de jouer un rôle central mais peut disposer d'atouts nouveaux, puisque les ambitions du Web Sémantique englobent les siennes. Les briques technologiques étant définies, il est permis de tout imaginer, d'actualiser toutes les applications en tenant compte de ce nouveau potentiel : fouille du web ou de réseaux d'entreprise, gestion de connaissances réparties, meilleure adaptation des résultats mais surtout des outils et des interfaces à des types d'utilisateur, recherche d'informations ciblées, extraction d'informations précises, combinaison de services web s'appuyant sur des connaissances, etc. Les outils de recherche d'information doivent donc s'adapter pour pouvoir traiter la couche « sémantique » prévue et en tirer le meilleur parti. Pour cela, ils doivent intégrer les technologies du web sémantique, largement inspirées des résultats de l'IA et de l'IC. Tout comme la plupart des applications ciblées dans le Web Sémantique, les logiciels de RI ont besoin d'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.

Or il s'agit là de sujets d'étude de la recherche d'information : (i) dépasser les limites des recherches basées sur des requêtes formulées par des mots-clés et sur des index constitués de sacs de mots tronqués et pondérés ; pour cela, (ii) s'appuyer sur des « concepts » qui rendent compte d'unités de « sens » exprimées dans les textes sous différentes formes linguistiques et sur lesquelles des inférences sont possibles [HAA 01]. En faisant appel à des concepts, on espère résoudre les problèmes de polysémie et d'ambiguïté des mots-clés, ou encore mieux identifier la signification des termes composés. De plus, les relations sémantiques entre concepts sont exploitées pour rapprocher des informations formulées autrement ou implicites, ou encore pour écarter des informations non souhaitées. Pour cela, la recherche d'information s'appuie sur des ressources comme les thésaurus fournissant des vocabulaires contrôlés dans lesquels des termes sont structurés. Avec le Web Sémantique, d'autres pistes ont été ouvertes, dont celle de l'utilisation de réseaux conceptuels ou de réseaux lexicaux, un des plus populaires étant WordNet, qui est en fait une base de données lexicales [MIL 95].

Or, parmi les moyens affichés par le projet du Web Sémantique [LAU 04], une des priorités est que ces réseaux conceptuels constituent des ressources partagées et réutilisables. L'hypothèse est que ces ressources seront d'autant plus pertinentes, qu'elles seront structurées et formalisées, définissant les notions-clés et la terminologie d'un domaine et permettant de raisonner sur ces connaissances. Ce sont les propriétés attendues des *ontologies*. La convergence entre les besoins de la recherche d'information et les propositions du Web Sémantique a été immédiate, ouvrant des perspectives prometteuses. Au-delà de la définition relativement restrictive de la notion d'ontologie proposée par le W3C, les travaux issus de ce

courant exploitent une grande diversité de modèles. Ces travaux, de plus en plus nombreux, couvrent les applications habituelles de la recherche d'information. Sans en dresser un panorama exhaustif, nous nous proposons dans cet article d'en faire ressortir les originalités et les avancées qu'ils ont permis de réaliser.

Au cœur de l'utilisation des ontologies pour la RI, se situe le processus consistant à associer des concepts à du contenu informationnel pour le représenter, le caractériser. Ce processus, appelé indexation conceptuelle ou annotation sémantique suivant qu'il soit automatisé ou manuel, fait appel au Traitement Automatique du Langage (TAL), et de plus en plus à l'apprentissage automatique et à l'extraction d'information [AMA 07]. Dans cet objectif, la richesse terminologique de l'ontologie exploitée est un atout pour mieux associer des concepts aux expressions linguistiques présentes dans les textes. Nous défendons donc l'idée que les ontologies sont des représentations des connaissances d'autant plus pertinentes pour la recherche d'information qu'elles comportent une dimension terminologique, ou même des éléments linguistiques permettant d'assurer le repérage de concepts et de relations sémantiques dans les textes.

1.3. Les ontologies

1.3.1 Héritage pluridisciplinaire des ontologies

La notion d'ontologie en tant qu'objet technique connaît aujourd'hui un succès exceptionnel en informatique. C'est certainement en partie parce qu'elle est l'aboutissement de l'évolution de modèles de connaissances issus de domaines scientifiques très différents. Mais c'est aussi parce que ce terme renvoie à un spectre très large de structures de données, caractérisées tantôt par leur format, tantôt par leur contenu. Les chercheurs du Web Sémantique insistent sur l'héritage philosophique (champ de l'Ontologie) et logique (ontologies formelles) [LAU 07]. Ce courant fait des concepts de l'ontologie de futurs prédicats logiques permettant de raisonner, essentiellement par classification. D'autres travaux situent 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]. En IC, 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 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 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.

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 [BOU04]. 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, bases de connaissances terminologiques [AUS 01], 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 (fig. 2). Or d'autres différences concernent la nature de leur contenu et justifient de revenir sur ce que sont ces structures.

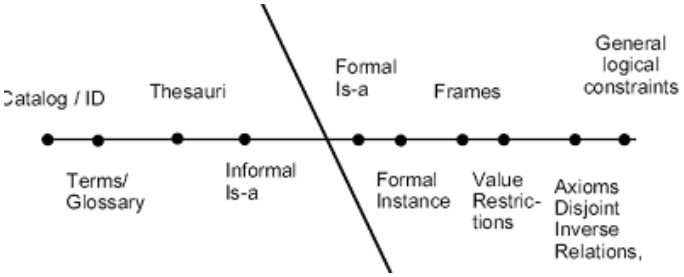

Figure 2 *: Différentes ressources terminologiquse et ontologies selon leur degré de formalisation[LAS 01]*

1.3.2 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. Les principaux types de langages documentaires sont les classifications et les thesaurus. Une *classification* répartit systématiquement en classes, des termes désignant des êtres, choses ou notions ayant des caractères communs 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 [REC 98].

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. L'utilisation des ontologies pour une indexation sémantique renouvelle encore la notion d'index. Un index « sémantique » se rapproche de ceux des documentalistes car il fait appel à une ressource normalisée pour caractériser un contenu informationnel. Pour comprendre les enjeux de ce basculement, revenons à la définition d'ontologie en IC.

1.3.3 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. 3). 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 : $H^C \acute{I} C \times C$, $H^C(C_1, C_2)$ signifie que C_1 est un sous-concept de C_2 (relation orientée)

Q Rel : **relation** $rel: R \to C \times C$ (définit des relations sémantiques non taxonomiques) avec 2 fonctions associées

dom : $R \to C$ avec dom(R):= $O1(rel(R))$

range : $R \to C$ avec range(R):= $\tilde{O}2(rel(R))$ 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 3 : *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. Ils sont identifiés et définies par les théoriciens d'un domaine, en dehors du contexte particulier dans

lequel on s'y intéresse [GUA 00]. Une approche plus pragmatique consiste à définir les concepts à partir de la manière dont une communauté, scientifique ou technique, les évoque à travers la langue utilisée. Ces ontologies devant être intégrées au sein d'applications ciblées, l'application et la tâche à réaliser sont prises en compte pour orienter les définitions. 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]. 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.

1.3.4 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 [REY 07]. En tant que discipline, la terminologie a réfléchi sur ses productions et sur l'articulation terme-concept. Ses conclusions apportent un éclairage sur la notion de concept et son statut par rapport au terme et à ses usages. Il nous paraît fondamental d'identifier en tant que telle la composante lexicale associée à une ontologie, selon la définition proposée dans [MAE 02a] (fig. 4). Une ontologie à composante lexicale est alors un couple (O, L) où O est une ontologie et L son lexique.

Le lexique d'une structure d'ontologie $O:= \{C, R, H^C, rel, AO\}$ 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 *:* $F(L) = \{C \text{ de } C / (L,C) \text{ estdans } F\}$ *et*
$F^{-1}(L) = \{L \text{ de } L / (L,C) \text{ est dans } F\}$
Idem pour G *et* G^{-1}

Figure 4 : *Définition du lexique d'une ontologie dans [MAE 02a]*

Des réflexions interdisciplinaires ont conduit à répercuter sur les ontologies les résultats établis pour les bases de connaissances terminologiques, 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] [REY 07]. 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.

1.3.5 Construire et réutiliser des ontologies

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]. Les ressources réutilisées peuvent aussi être des thésaurus ou des terminologies [VEL 05], c'est-à-dire des ressources qui étaient utilisées jusque là en recherche d'information pour former des index plus pertinents dans des domaines spécialisés.

L'utilisation de textes comme sources de connaissances connaît un regain depuis 2000. Il s'explique en partie par des avancées du TAL, qui ont permis de disposer non seulement d'études statistiques des textes, mais de ressources et d'outils permettant leur analyse selon des méthodes linguistiques robustes. Des logiciels spécialisés 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, OntoLT, 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. Depuis 2003, on assiste à une véritable explosion de l'exploitation des textes, avec deux tendances fortes : l'utilisation du web comme corpus et la recherche systématique de l'automatisation, en particulier par des techniques d'apprentissage et d'extraction d'information [CIM 05]. On parle ainsi d'*ontology learning* [MAD 02] [BUI 05]. L'apprentissage associé à des techniques d'extraction d'information et d'annotation sert alors à retrouver dans des textes des traces linguistiques de la présence de concepts ou d'instances de concepts, ce qui permet de compléter l'ontologie par le processus appelé *ontology population*. Nous reviendrons plus tard sur ces techniques car, à travers le repérage d'instances de concepts, elles conduisent à annoter automatiquement ou à indexer des textes à l'aide de ces instances de concepts [AMA 06] [BUI 05].

Pour la recherche d'information, les ontologies construites à partir de textes présentent plusieurs intérêts, sur lequel nous reviendrons. Tout d'abord, elles comportent souvent une composante terminologique plus riche que dans celles construites manuellement, puisque les textes donnent accès à diverses formulations

des concepts. Or cette composante est primordiale pour une bonne indexation par les concepts [ROU 02]. Ensuite, les concepts qu'elles contiennent sont un bon reflet de ceux contenus dans les textes qui ont servi à les construire. Enfin, les outils qui ont servi à les construire peuvent ensuite servir, de manière symétrique, à retrouver les concepts ou leurs instances dans de nouveaux textes similaires, dans un objectif d'annotation [AMA 07].

Un des points importants pour la qualité de l'ontologie, et qui fait sa différence avec un thésaurus ou un réseau sémantique, est la qualité de la définition des concepts qui la composent, ainsi que l'explicitation de choix de structuration, relatifs à ce qui est appelé un « engagement ontologique ». La définition des concepts passe par leur situation dans une hiérarchie et la mise en forme de relations avec d'autres concepts ou de propriétés. Cet effort de structuration a un coût, celui de définir des classes de haut niveau pertinentes, puisqu'elles vont avoir une incidence (en cascade) sur la définition des classes plus spécifiques au domaine. Les parties hautes des ontologies (top-ontologies) DOLCE [MAS 03], SUMO[3] ou celles proposées par des chercheurs comme Brachman ou Sowa sont autant de points de vue sur les notions primitives qui permettent de penser le monde. Or il paraît souvent plus simple, dans un domaine donné, de d'intéresser directement aux concepts de ce domaine. Le risque est alors de limiter la compréhension par d'autres personnes ou systèmes, le manque de réutilisabilité ou la difficulté à mettre en correspondance de telles ontologies.

Pour mettre en avant les engagements ontologiques, plusieurs méthodes proposent de rendre explicites des critères de normalisation de l'information tirée des textes. Dans la lignée des travaux d'Aristote, la méthode ARCHONTE [BAC 04] propose de formuler des critères de différenciation entre concepts, qui guident leur organisation. La méthode TERMINAE reprend ces principes [AUS 05a] [SZU 02]. Dans ONTOCLEAN, Guarino et Welty [GUA 00] proposent des méta-propriétés permettant de caractériser les concepts, et de vérifier que leur mise en relation dans l'ontologie ne conduit pas à des incohérences sémantiques. La méthode ONTOSPEC [KAS 02] permet aussi de mettre en forme de tels critères en s'appuyant sur les méta-propriétés d'ONTOCLEAN. La légitimité des définitions de concepts peut donc renvoyer soit à l'expression des connaissances dans la langue soit aux méta-propriétés vérifiées par les concepts et les relations.

Aujourd'hui, suite aux efforts de normalisation du W3C, de nombreuses « ontologies » sont disponibles sur le web dans le format standard OWL, afin d'en faciliter la réutilisation. Ces modèles vont d'ontologies génériques pour de grands

[3] http://www.ontologyportal.org/index.html (nov. 2007), SUMO : Suggested Upper Merged Ontology

domaines scientifiques ou techniques (par exemple, LKIF-Core[4] se veut un noyau générique pour le domaine du droit [BRE 05], le CRM[5] de CIDOC un cadre général pour décrire des objets d'art et patrimoniaux [DOE 03]), des ressources proches de la langue comme la base de données lexicale WordNet [MIL 95], mais surtout des milliers de modèles très spécialisés, plus ou moins riches, et pour lesquels est rarement explicité l'engagement ontologique qui a guidé leur construction. Un outil de recherche en ligne, SWOOGLE[6], est d'ailleurs disponible pour retrouver « rapidement » des ontologies écrites en OWL sur des sujets particuliers.

1.4. Utilisation des ontologies en recherche d'information

1.4.1 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 00a].

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 qualifier le contenu de bases de données, de retrouver les documents, parties de documents ou données 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 et des données 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.

---

[4] Ontologie LKIF-Core : http://www.estrellaproject.org/lkif-core/ (nov. 2007)

[5] CIDOC – CRM ontologie RDFs : http://cidoc.ics.forth.gr/rdfs/Cidoc_v4.2.rdfs (nov. 2007)

[6] http://swoogle.umbc.edu/ : en nov. 2007, le site propose de rechercher parmi 10 000 ontologies.

Dans le cadre du Web Sémantique, la recherche d'information se retrouve ainsi renouvelée sous plusieurs points de vue [LAU 07]. Elle se trouve face à des ressources multiples parmi lesquelles l'information prend des formes très variées, et est disponible en très grande quantité. Les questions soulevées sont à la fois la prise en compte de l'échelle du web, donc la capacité à traiter de très gros volume de sources d'information, et l'intégration de la diversité des formats, de leur hétérogénéité, dans différents types d'applications. Ces applications couvrent la fouille du web à la recherche d'informations précises, la sélection de sous-ensembles d'information ou de documents répondant à des requêtes, l'exploitation de données structurées ou semi-structurées, la combinaison d'informations provenant de différentes sources, leur intégration sémantique, mais aussi la définition de « méta-moteurs » ou de services web capables de combiner plusieurs types de recherches pour rendre des services élaborés.

1.4.2 Apports attendus des ontologies dans la recherche d'information

Au cœur du projet du Web Sémantique [CHA 04b], les ontologies sont considérées comme le moyen de disposer de modèles de connaissances partageables, consensuels et permettant de raisonner sur les connaissances. Les ontologies peuvent alors contribuer à différents aspects de la recherche d'information, et être vues sous différents angles, que nous détaillerons par la suite.

- ressources fournissant des méta-données : considérées comme une version élaborée et formelle des thésaurus et langages documentaires, leur formalisation permet d'élargir les possibilités de caractérisation des documents et des besoins en information des utilisateurs. Elles sont alors le creuset de mots-clés servant à définir des méta-données, à caractériser le contenu informationnel des documents ou à les indexer. Elles peuvent être exploitées pour une indexation manuelle ou automatique, des documents et des requêtes, appelée indexation conceptuelle ou sémantique. L'intérêt d'un index composé d'éléments d'ontologie est qu'il se détache de la langue, qu'il est exprimé dans un langage se prêtant à des traitements informatiques, et qu'il est possible de faire référence à des ontologies partagées pour mieux regrouper ensuite des collections annotées. La difficulté à élaborer automatiquement ces index découle de l'ambiguïté de la langue, la polysémie des termes ou encore le complexité, qui rendent difficile la reconnaissance des bons concepts. Enfin, les ontologies servent à mieux restituer les documents résultats, que l'on peut mieux classer et ordonner.

- langages partagés et formats d'échange de données : les ontologies servent aussi à redéfinir les modes d'interrogation de sources de données hétérogènes, en décrivant à un niveau conceptuel ces données ou les schémas de bases de données [ROU 03].L'ontologie sert alors de langage pivot, qui permet aux utilisateurs de

formuler les requêtes de manière uniforme, sans avoir à se préoccuper de la disparité des modèles de données de chacune des sources consultées. Cette approche suppose de développer des traducteurs (wrappers) capables de traduire la requête formulée à l'aide de concepts de l'ontologie vers chacun des langages et modèle conceptuel des ressources à consulter. En extraction d'information, elles permettent de décrire formellement et sémantiquement les informations recherchées en se détachant de leur formulation dans une langue particulière [NED 05]. Les ontologies sont aussi utilisées pour faciliter les échanges entre des agents logiciels effectuant des recherches complémentaires sur des sources de forme et contenu différents

- structures de données permettant de capitaliser des données, de caractériser des données informelles pour mieux les rechercher ensuite. L'ontologie est alors intéressante comme mode de stockage de ces données et connaissances.

Dans cet article, nous nous focalisons principalement sur le premier aspect.

1.5. Ontologies, indexation et annotation sémantique

1.5.1 Ontologies pour décrire le contenu de documents

Tout d'abord, la recherche d'information s'est tournée 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.).

1.5.2 Indexation conceptuelle versus annotation

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] [AMA 07]. 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[7]. 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. A titre d'exemple, le texte de la figure 5 serait indexé de manière classique par une liste de mots tronqués et pondérés (figure 6a, sur laquelle les poids n'ont pas été représentés) et pourrait être indexé à l'aide d'instances de concepts d'une ontologie décrivant des conférences (figure 6b). Une annotation sémantique à l'aide de concepts OWL consisterait à associer quelques instances de concepts plus riches et plus complets pour décrire la même information (figure 6c).

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

3rd European Semantic Web Conference (ESWC2006)

The 3rd Annual European Semantic Web Conference (ESWC2006) will be held in Budva, Montenegro from the 11th - 14th June, 2006.

It will present the latest results in research and application in Semantic Web technologies (including knowledge markup languages, Semantic Web services, ontology management and more).

ESWC 2006 will also feature a special industry-oriented event providing European industry with an opportunity to become even more familiar with these technologies. It will offer a tutorial program, focusing on the latest in Semantic Web technologies.

Figure 5 : *Texte original à indexer (tiré du site http://www.eswc2006.org)*

Europe	European [geographic area: "Europe"]
Semanti	
Web	Semantic Web [semantic Web]
Confere	Conference [conference: conference1]
Annual	Annual [time-frequency: "annual"]
Europe	European [geographic area: "Europe"]
Confere	Conference [conference]
Budva	Budva [City: "Budva"]
Montene	Montenegro [geographic area: "Montenegro"]
present	present [to present]
results	results [result]
Researc	Research [research]
Applica	Application [application: application1]
Technol	Technology [technology]
Includi	
Knowled	
Markup	Knowledge Markup [Knowledge Markup : KM1]

Figures 6a et 6b : *Index classique et index sémantique possible du texte de la figure 5. Les mots entre [] correspondent à des concepts (comme* `research`*) ou à des instances de concepts(comme* `geographic area: «Europe»`*) d'une ontologie*

```
<eswc:ConferenceEvent rdf:about="http://www.eswc2006.org/#">
  <dc:title>3rd European Semantic Web Conference (ESWC2006)</dc:title>
  <dc:description>The 3rd Annual European Semantic Web Conference
(ESWC2006) will be held in Budva, Montenegro from the 11th – 14th June,
2006. It will present ….</dc:description>
  <eswc:hasLocation rdf:resource="http://www.eswc2006.org/places/#Maestral"
/>
  <eswc:hasStartDateTime>2006-06-11T09:00:00+02:00</eswc:hasStartDateTime>
  <eswc:hasEndDateTime>2006-06-14T17:45:00+02:00 </eswc:hasEndDateTime>
  <eswc:hasProgramme
rdf:resource="http://www.eswc2006.org/conference_programme.html" />
  <eswc:hasProceedings
rdf:resource="http://www.springer.com/uk/home/computer/database?SGWID=3-
40109-22-173660166-0" />
```

Figures 6c : *Instance du concept de* `ConferenceEvent` *annotant le texte de la figure 5. Ce concept est tiré de l'ontologie* `eswc` *et possède les propriétés indiquées par les mots en gras (*`description, hasLocation, hasprogramme …`*)*

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 [PRI 04]. Pour repérer des instances de concepts dans des 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 05c]. Les mêmes recherches permettent donc de progresser sur ces deux fronts, et constituent une des clés actuelles de la réussite du Web Sémantique. Les coûts de la construction de l'ontologie et de l'indexation sont encore élevés. 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.

1.5.3 Ontologie pour l'indexation sémantique en amont de moteurs généralistes

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. Les moteurs n'étant pas prévus pour manipuler des représentations sémantiques, des étapes de conversion sont nécessaires pour exploiter au mieux les concepts tout en revenant à une représentation comprise par un moteur classique [MAY 03]. 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 servant à indexer 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. La question n'est donc pas de savoir si les ontologies améliorent la recherche d'information mais plutôt dans quelles conditions les ontologies peuvent l'améliorer : pour quelles tâches de RI, quel type d'ontologies sont les mieux adaptées (contenu, degré de formalisation, couverture du

domaine), quelles heuristiques optimales permettent d'exploiter au mieux les relations entre concepts.

La représentation sémantique des documents proposée par M. Baziz [BAZ 05b], 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 et seuls les concepts d'un poids supérieur à un seuil sont retenus. 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. 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.

1.5.4 Extraction d'information pour l'annotation sémantique

Dans plusieurs domaines spécialisés comme la médecine, la biologie et depuis quelques années la génétique, l'annotation sémantique connaît un succès croissant afin de recouper les connaissances dispersées au sein de grandes collections de textes scientifiques [NED 05]. Ces domaines présentent l'avantage de disposer de ressources telles que des thésaurus, des inventaires de noms de gênes et des

ontologies médicales ou pharmacologiques. Il est désormais classique d'envisager des chaînes de traitement automatique du contenu de ces textes de manière à retrouver des informations précises et structurées, qu'il est aisé ensuite de stocker dans des bases de données et de recouper pour « découvrir » des connaissances nouvelles. Dans l'approche développée par F. Amardeilh [AMA 07], des connaissances structurées sont utilisées dans des règles d'extraction, et mise en correspondance avec des concepts d'une ontologie, de manière à organiser les connaissances extraites à l'aide de ces concepts. La recherche d'information s'appuie alors sur les données extraites, qui indexent les documents dont elles sont tirées. Des applications dans différents domaines ont montré l'efficacité de l'approche dans des domaines thématiquement bien délimités, mais sur de gros volumes de documents (articles de loi, presse grand public, etc.).

Dans le projet de [KAM 07], l'idée est d'aider au ciblage des gènes responsables dans les maladies génétiques, et ce par le biais des localisations chromosomiques identifiées au cours d'expériences rapportées dans des articles scientifiques. Plusieurs critères bibliographiques comme l'expression tissulaire, les résultats des puces à ADN, les fonctions déjà connues peuvent alors permettre de sélectionner les candidats les plus pertinents. Par contre, l'information concernant les anomalies génétiques détectées, localisées et liées potentiellement à un phénotype ne sont présentes que dans des articles accessibles par des Bases de Données Documentaires (PubMed, Medline, OMIM). Une chaîne de traitement automatique du contenu des textes a été développée pour extraire de ces documents toute l'information pertinente destinée à enrichir les bases de données. Il est à noter que cette information découle généralement d'une expérience, basée sur un nombre de patients, pouvant présenter des similarités d'anomalies chromosomiques. L'annotation sémantique se fera donc à l'aide de concepts identifiés dans les textes grâce à des règles d'annotation. Ces concepts doivent à la fois décrire les chromosomes, leur localisation et les maladies repérées ainsi que le contexte de leur observation. Les concepts d'échantillon, de conditions de l'expérience et de résultat sont autant d'éléments de contexte représentés par les concepts suivants (en gras) reliés à d'autres concepts de l'ontologie (concepts et relations correspondent aux mots soulignés) :

- **Echantillon** (données de l'expérience) :
  - *nombre* d'*individus* atteints d'une même *pathologie*
  - *nombre de familles ou pedigrees*
  - *nombre de lignées cellulaires provenant d'individus atteints d'une même pathologie*
  - *masse de données provenant d'individus atteints d'une même pathologie*
- **Conditions de l'Expérience** :
  - *type d'analyse* effectuée sur un Echantillon
- **Résultat** :
  - *nombre* de *patients* (ou *fréquence*) porteurs d'une *anomalie*

chromosomique localisée dans une *région*
- *région* transmise associée à un *LOD Score* ou *NLP Score*

1.6. Ontologie pour formuler des requêtes, sélectionner et visualiser les documents résultats

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].

1.6.1 Formulation de requêtes à l'aide de concepts

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.

Expansion systématique des requêtes : Le système OWLIR [MAY 03] suggère une solution possible pour utiliser une indexation à l'aide d'ontologie, qui consiste à combiner la recherche d'information classique à une recherche sémantique. Comme montré sur la figure 7, le processus qui traite une requête consiste à n'utiliser la sémantique que pour reformuler les requêtes en terme de concepts (ou de méta-données) puis à les reformuler sous forme d'un ensemble plus riche de termes. Ensuite, les connaissances sont prises en compte pour filtrer les résultats d'un moteur généraliste. On peut espérer ainsi améliorer au mieux la qualité des premiers résultats présentés, mais le risque est grand d'introduire du bruit (documents non pertinents retrouvés par le moteur classique). C'est pourquoi les auteurs ont complété ce processus par la prise en compte en entrée de connaissances selon deux biais : en premier lieu, au moment de la reformulation de la requête, des concepts identifiés sur des textes jugés pertinents lors de précédentes interrogations guident la

reformulation de la requête ; puis des textes pertinents sont fournis au moteur de recherche pour orienter la recherche de résultats analogues. Ces textes seront enrichis de termes tirés de l'ontologie servant à annoter, ce qui remplace, pour ces auteurs, l'indexation conceptuelle.

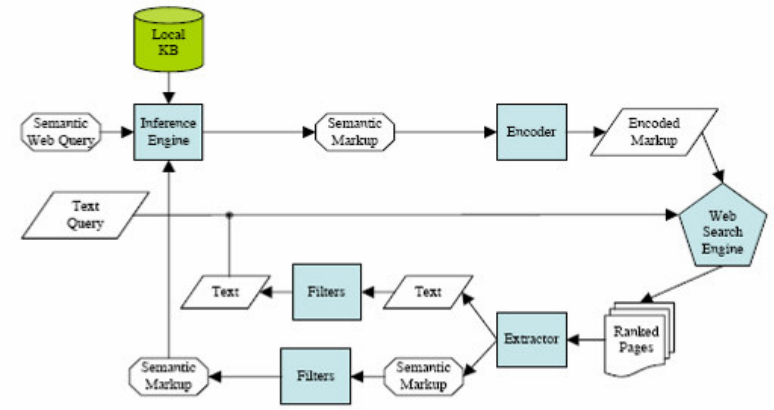

Figure 7*: Le processus de recherche d'information de OLWRI [MAY 03]*

Importance de la nature des relations dans l'expansion de requêtes : Bien que séduisante, l'expansion de requêtes conduit à prendre le risque de générer du bruit, en introduisant des termes inadaptés qui retourneront des documents non pertinents pour l'utilisateur. L'idée de M. Baziz a été d'assurer une amélioration systématique des résultats en procédant à une « expansion prudente » à l'aide de la base de données lexicales Wordnet. 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 (réalisés avec le moteur Mercure et la collection Clef2001) ont montré que la nature des relations sémantiques a une influence significative sur l'expansion de requête [BAZ 03]. De plus, ce processus ne conduit à une amélioration significative de la pertinence des réponses fournies par le moteur que si l'on minimise le nombre de concepts utilisés pour représenter une requête, et si l'on n'exploite que les relations est-un pour l'expansion (les relations de méronymie ou d'antinomie, au contraire, détériorent les résultats).

1.6.2 De la requête aux résultats dans le cas d'une indexation sémantique

Une fois la requête et les documents indexés par des concepts, l'appariement se fait en évaluant la proximité entre la représentation de la requête et celle correspondant à chaque document. Le calcul de similarité dépend de la

représentation choisie et fait appel le plus souvent à 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.

Ontologies pour l'exploration documentaire : OntoExplo est un environnement d'exploration de collections documentaires à partir de hiérarchies de concepts de ce domaine et de concepts décrivant la tâche de veille [HER 05b]. 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 interroger la collection de documents en naviguant dans cet espace d'information. 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.

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. La présentation de hiérarchies à l'utilisateur permet de simplifier la représentation du domaine qui lui est montrée. Ces hiérarchies sont pourtant unifiées en une ontologie, trop complexe pour être présentée directement lors de l'exploration de collections. La figure 4 présente un exemple d'exploration de corpus d'articles 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).

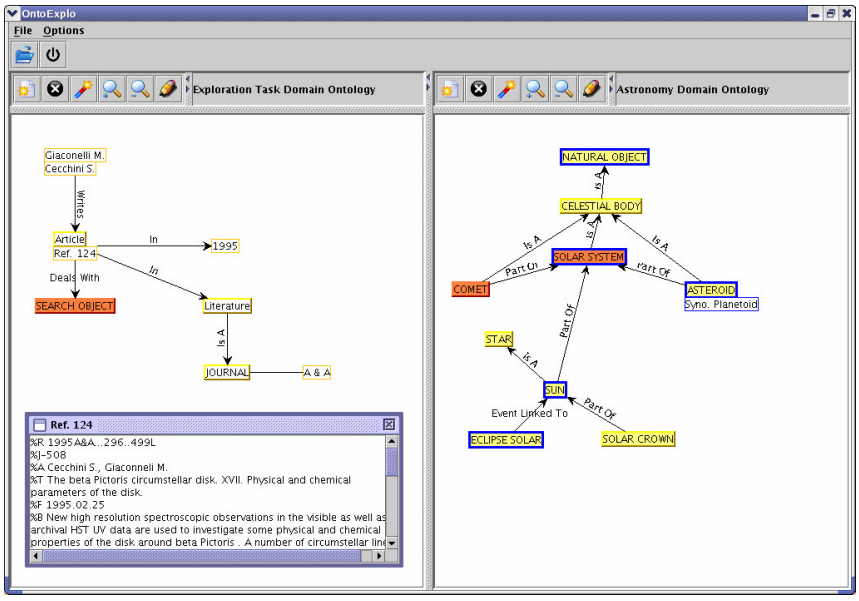

Figure 4 : *Visualisation de la connaissance établie pour un article à partir du corpus*

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.

1.6.3 Les raisonnements permis par les ontologies au cours de la RI

Un des éléments mis en avant dans l'usage d'ontologie est la capacité de raisonner en exploitant les relations hiérarchiques et les autres relations sémantiques entre concepts. Or ce type de raisonnement n'est pas trivial. Nous avons mentionné la nécessité d'ne faire usage avec prudence pour l'expansion de requête, comme l'a montré M. Baziz. Nus donnons ici deux exemples permis par les raisonnements possibles au sein d'une ontologie formalisée.

Dans le système PICSEL [ROU 03], l'ontologie sert de langage pivot pour interroger plusieurs sites de formats et de contenus hétérogènes dans le domaine du tourisme. Ce type de système simplifie les interactions avec l'utilisateur, mais il se heurte aux difficultés classiques de l'interrogation de bases de données. Ainsi, certaines requêtes échouent car aucune réponse ne leur correspond ; ces échecs peuvent venir d'une mauvaise connaissance du domaine par les utilisateurs (qui demandent des données qui n'existent pas, par exemple un séjour à la plage à Madeire où il n'y a pas de plage) ou bien de lacunes dans les ressources interrogées. Dans ce cas, les utilisateurs aimeraient être guidés pour savoir pourquoi la requête a échoué, et comment la reformuler pour obtenir une réponse proche. Par ailleurs, la formulation des requêtes peut être lourde et les utilisateurs ont envie de conserver et combiner des éléments de requêtes déjà formulées. Ainsi, les auteurs de PICSEL montrent dans [ROU 03] tout le potentiel de la formalisation de l'ontologie à l'aide d'un langage logique. La formalisation permet ainsi de définir des concepts complexes faisant intervenir de nombreux concepts élémentaires (sous forme de formuler logique, comme la notion de vol aller-retour ou de plusieurs formules, ce qui revient à les définir par une disjonction) et de pouvoir les utiliser dans des requêtes pour en simplifier la formulation. La formalisation permet de calculer et fixer des enchaînements de relations entre concepts, toujours pour les réutiliser. Enfin, elle permet d'affiner des requêtes et de calculer des plans de requêtes. Par exemple, pour affiner des requêtes en échec, le système repère la source des conflits parmi les concepts de la requête, exploite la hiérarchie des concepts pour généraliser celui qui pose problème, puis calcule des requêtes satisfiables à partir des requêtes insatisfiables en exploitation la généralisation de concepts.

Le raisonnement intervient également dans le système oMAP pour reformuler des requêtes en s'appuyant sur l'alignement d'ontologies (ontology mapping) [STR

06]. Dans oMAP, un agent de recherche procède en trois étapes : il sélectionne des sources d'information pertinentes (pour cela, au lieu de balayer tout le web le système s'appuie sur l'ontologie pour filtrer des ressources) ; ensuite, il reformule la requête de l'utilisateur en fonction d'une ontologie de référence, de manière à multiplier les formulations possibles de la requête et à trouver celles présentes dans les documents ; enfin, il combine les différents résultats (cela revient à aggréger et fusionner les textes pertinents au sein d'une liste unique et classée). A chaque étape, un classifieur cherche à reconnaître le meilleur concept associé aux termes présents dans un texte et dans une requête, pour la reformuler et mieux gérer l'association texte-requête. Il raisonne sur la place des concepts dans la hiérarchie de concepts et sur leurs relations pour prévoir un concept plus générique ou plus précis.

1.7. Quel web sémantique pour une recherche d'information efficace ?

1.7.1 Adéquation des ressources sémantiques à la recherche d'information

A travers ces analyses, nous avons voulu 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 ressource 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 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.

1.7.2 Annotation sémantique et maintenance d'ontologie

Parmi les perspectives de recherche qui se dégagent, les sujets d'actualité concernent l'automatisation coordonnée et continue du processus de construction d'ontologie et du processus d'indexation sémantique. Nous avons souligné que les mêmes techniques de fouille de textes, faisant appel à des algorithmes de TAL, d'extraction d'information et d'apprentissage, pouvaient servir dans ces deux cas. 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. Mais surtout, le contexte de l'indexation sémantique soulève de manière cruciale la question de la maintenance des ontologies. Dans de nombreux contextes applicatifs, les collections documentaires et les besoins des utilisateurs évoluent régulièrement (nouveaux articles scientifiques, dépêches de presse, articles de loi à indexer) et, avec elles, les connaissances et le vocabulaire du domaine concerné. Grâce à l'apprentissage automatique, plusieurs travaux envisagent un processus cyclique associant annotation et enrichissement de l'ontologie (pour définir de nouveaux concepts ou de nouvelles instances). Cette approche est à l'origine des choix du projet Dynamo, où un système multi-agents représente l'ontologie et assure cette maintenance. Elle justifie également d'associer étroitement enrichissement d'ontologie et annotation dans OntoPop [AMA 07] ou TextViz [REY 07].

1.7.3 Complémentarité du web sémantique et du web 2.0

Présentés il y a quelques années comme deux alternatives opposées, le web sémantique et les technologies favorisant la prise en charge par les utilisateurs du contenu du web (web 2.0) semblent aujourd'hui complémentaires. Mieux, exploiter cette complémentarité pourrait être une solution prometteuse pour dépasser certaines limites du web sémantique [MAY 03]. Par exemple, les données et connaissances diffusés sur le web avec les technologies du web 2.0 simplifient les tâches de recherche d'information. En favorisant des modes d'accès à l'information basés sur le classement et l'annotation par l'utilisateur, sur des systèmes qui l'alertent de la présence d'information nouvelle ou mise à jour avant qu'il n'aille la chercher, on

réduit les recherches basées sur des requêtes formulées en langage naturel sur des moteurs généraux. L'objectif est bien sûr de faire évoluer la recherche d'information classique pour mieux exploiter les pages déjà indexées par des méta-données issues de folksonomies et non d'ontologies formelles, ou, mieux, de savoir gérer les deux types d'annotation de manière transparente pour les utilisateurs. Cela signifie savoir lire ces annotations, mais surtout savoir les exploiter alors qu'elles sont « mal définies » et qu'il est encore plus risqué de s'appuyer sur leur structure. Le prolème est également celui du format des annotations, qui ne sont plus homogènes (OWL d'un côté, html de l'autre). Enfin, les communautés d'intérêt mettent à jour régulièrement des ressources annotées et des thésaurus d'annotation qu'il peut être intéressant d'exploiter soit pour accéder directement à des documents, soir pour constituer des ontologies.

1.7.4 Ontologies ou thésaurus structurés pour la recherche d'information ?

Enfin, l'utilisation des ontologies et des terminologies en recherche d'information vient alimenter deux débats : l'un sur ce que doivent être les ontologies, sur la faisabilité de définir des ressources stables et partageables sur le web ; l'autre sur ce que doivent être les ressources pertinentes pour la recherche d'information. Nous insistons ici sur ce dernier point, car plusieurs chercheurs s'interrogent sur le degré de formalisation utile pour un modèle de connaissances qui faciliterait la recherche d'information. Ainsi, une communauté scientifique fait le pari que des thésaurus structurés sont suffisants pour répondre au besoin d'indexation sémantique, d'expansion de requête et de représentation riche de l'articulation entre termes et concepts. Cette communauté a défini, en accord avec le W3C, le standard SKOS[8] (Simple Knowledge Organisation Systems) pour la représentation des thésaurus structurés. SKOS est une surcouche de RDFs relativement compatible avec OWL. Sa capacité d'expression est beaucoup moins riche, mais ce langage présente l'avantage de simplifier la représentation des connaissances et d'alléger le volume des modèles construits. Les applications utilisant SKOS commencent à voir le jour, en particulier pour l'annotation de documents multi-média. Cette tendance converge avec celle de l'influence du Web 2.0 , et laisse présager une grande variété de réalisations et d'expériences pour la recherche d'information dans le cadre du web sémantique.

1.8. Bibliographie

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

8 http://www.w3.org/2004/02/skos/

[AUS 01] AUSSENAC-GILLES N., CONDAMINES A., Entre textes et ontologies formelles : les bases de connaissances terminologiques. in *Ingénierie et capitalisation des connais-sances*. 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. in *RIAO 2004*, 129-142. 2004.

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

[AUS 05a] 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. 40-72. 2005.

[AUS 05b] 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 05c] 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

[AMA 06] AMARDEILH F., OntoPop or how to annotate documents and populate ontologies from texts, In *Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, Budva (Montenegro), CEUR Workshop Proceedings, ISSN 1613-0073, 2006.

[AMA 07] AMARDEILH F., *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. Mémoire de thèse de l'université Paris X. Mai 2007

[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.

[BAG 04] BAGET J.F., CANAUD E., EUZENAT J., SAÏD-HACID M., Les langages du Web sémantique in CHARLET J., LAUBLET P., REYNAUD C., (Eds.) Numéro spécial « Web sémantique », *Revue Information, Interaction, Intelligence I3*, Cépadues-Editions, 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, **8** (4) 2003, 113-136.

[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), F. Crestani and I. Ruthven (Eds.): LNCS 3507. Berlin : Springer-Verlag, 171-186, 2005.

[BER 98] BERNERS-LEE T., FIELDING R., MASINTER L. (1998). Uniform Resource Identifiers (URI): Generic Syntax. Request for Comments 2396, IETF. http://www.ietf.org/rfc/rfc2396.txt

[BER 01a] BERNERS LEE T., HENDLER J., LASSILA O., The semantic Web. *Scientific American*, New York, May, 35-43. 2001.

[BER 01b] BERNERS LEE T., W3C Semantic Web Activity Statement. http://www.w3.org/2001/sw/Activity, 2001.

[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, Toulouse3. 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., MAGNINI B., *Ontology Learning from Text: Methods, Evaluation and Applications*. Volume 123, Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

[BRE 05] BREUKER J., VALENTE A., WINKELS, Use and Reuse of Legal Ontologies, *in Law and the Semantic Web*, Benjamins V. R., Casanovas P., Breuker J. \& Gangemi A. (ed.), Springer Verlag, 36--64, 2005.

[BRI 99] BRICKLEY Dan & GUHA R., Eds. Resource description framework schema specification. Proposed recommandation, W3C. http://www.w3.org/TR/PR-rdf-schema , 1999.

[BRI 02] BRICKLEY Dan & GUHA R., Eds., RDF Vocabulary description language 1.0: RDF Schema. Working draft, W3C. 2003. http://www.w3.org/rdf-schema

[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 IC 2005*, Lyon (F), 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 04a] CHARLET J., LAUBLET P., REYNAUD C., (Eds.) Numéro spécial « Web sémantique », *Revue Information, Interaction, Intelligence I3*, Cépadues-Editions, 2004.

[CHA 04b] CHARLET J., BACHIMONT B., TRONCY R., Ontologies pour le Web sémantique, in *Revue Information, Interaction, Intelligence I3*, Charlet J., Laublet P., Reynaud C., (Eds.) Numéro spécial « Web sémantique », Cépadues-Editions, 2004.

[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.

[DOE 03] DOER M., The CIDOC CRM, An ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*. **24** (3). 2003.

[EUZ 04] EUZENAT J., Chouette alors! Un langage pour les ontologies sur le web, Actes des *15e journées Francophones d'Ingénierie des Connaissances (IC 2004),* Grenoble, France, PUG, 2004. http://liris.cnrs.fr/~ic04/programme/Euzenat.pdf

[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.

[HAA 01] HAAV, H. M., LUBI, T.-L A Survey of Concept-based Information Retrieval Tools on the Web. In *Advances in Databases and Information Systems*, Proc. of 5th East-European Conference ADBIS*2001, A. Caplinkas and J. Eder, (Eds), Vol 2., Vilnius "Technika", 29-41. 2001.

[HER 05a] HERNANDEZ N., *Ontologies de domaine pour la modélisation du contexte en recherche d'information*. Thèse de doctorat l'université Paul Sabatier (Toulouse 3), 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.

[KAM 07] KAMEL M., PERRET E., Extraction d'Information pour le ciblage des gènes impliqués dans les maladies génétiques, in *Actes des Journées Ouvertes Biologie, Informatique et Mathématiques (JOIM 2007*), Marseille, 2007.

[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.

[KEL 04] KELLER P., TOUMANI F., Les Web services sémantiques, in LAUBLET P. , CHARLET J., REYNAUD C., Introduction au web sémantique, in *Revue Information, Interaction, Intelligence I3*, Charlet J., Laublet P., Reynaud C., (Eds.) Numéro spécial « Web sémantique », Cépadues-Editions, 2004.

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

[LAU 04], LAUBLET P. , CHARLET J., REYNAUD C., Introduction au web sémantique, in *Revue Information, Interaction, Intelligence I3*, Charlet J., Laublet P., Reynaud C., (Eds.) Numéro spécial « Web sémantique », Cépadues-Editions, 2004.

[LAU 07] LAUBLET P.**,** Web sémantique et ontologies, in *Humanités numériques. Nouvelles technologies cognitives et concepts des sciences sociales*, Brossaud C. et Reber B. (dir.) Hermès, 2007.

[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., SCHNEIDER L., *The WonderWeb Library of Foundational Ontologies and the DOLCE ontology*. WonderWeb Deliverable D18, Final Report (vr. 1.0, 31- 12-2003). 2003.

[MAY 03] Mayfield J., Finin T. Information retrieval on the Semantic Web:vIntegrating inference and retrieval, in *SIGIR 2003 Semantic Web Workshop*, Toronto, Canada. 2003.

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

[MIL 95] MILLER G., Wordnet: A lexical database. *Communication of the ACM*, **38**(11):39-41. 1995.

[MIL 06] MILES A., Retrieval and the Semantic Web, Oxford Brookes University, M.Sc Dissertation, Sept. 2006.

[MOR 99] MORIN E., « Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques », *Traitement Automatique des Langues*, **40**, Numéro 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 3. Nov. 2000.

[NED 05] NÉDELLEC N., NAZARENKO A., "Ontology and Information Extraction: A Necessary Symbiosis", in *Ontology Learning from Text: Methods, Evaluation and Applications*, P. Buitelaar, P. Cimiano and B. Magnini (Eds.) , IOS Press Publication: Amsterdam, 2005.

[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.

[PRI 04] PRIE Y., GARLATTI S., Méta-données et annotations dans le Web sémantique, in Charlet J., Laublet P., Reynaud C., (Eds.) Numéro spécial « Web sémantique », *Revue Information, Interaction, Intelligence I3*, Cépadues-Editions, 2004.

[REC 98] RECTOR A. L.. Thesauri and formal classifications: Terminologies for people and machines. *Methods of Information in Medicine*, **37**(4–5), 501–509. 1998.

[REY 07] REYMONET A., THOMAS J., AUSSENAC-GILLES N., Modélisation de Ressources Termino-Ontologiques en OWL. *18° journées francophones d'Ingénierie des Connaissances (IC 2007).* Grenoble (F). F. Trichet (Ed.), Cépaduès Editions, 169-180, 2007.

[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 02] ROUSSEY C., CALABRETTO S., PINON J.-M., Le thésaurus sémantique : contribution à l'ingénierie des connaissances documentaires. In B. Bachimont, Ed., *Actes des 6es Journées Ingénierie des Connaissances*, 209–219, Rouen (F.). 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.

[STR 06] STRACCIA U., TRONCY R., Towards Distributed Information Retrieval in the Semantic Web: query Reformulation Using the oMAP Framework. in Sure Y., Domingue J. (Eds.): Proceedings of ESWC 2006, Spriner Verlag, LNCS 4011, 378-392, 2006.

[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*. 1996.

[VEL 05] VELARDI, P., NAVIGLI, R., CUCHIARELLI, A., NERI, F., Evaluation of Ontolearn, a methodology for automatic population of domain ontologies. In BUITELAAR, P., CIMIANO, P., MAGNINI, B. (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press: Amsterdam. 2005.