Exploitation du web sémantique pour la veille technologique

Tuan Dung Cao

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS - UFR Sciences
Ecole Doctorale Sciences et Technologies de l'Information et de la Communication (S.T.I.C)

T H È S E

pour obtenir le titre de
Docteur en Sciences
de l'UNIVERSITE de Nice-Sophia Antipolis
Spécialité

Informatique

présentée et soutenue par

Tuan Dung CAO

Le 29 Novembre 2006

Exploitation du web sémantique pour la veille technologique

Thèse dirigée par ***Rose DIENG-KUNTZ***

Jury :

Président	Nhan Le Thanh
Rapporteurs	Chantal Reynaud
	Parisa Ghodous
Examinateurs	Rose Dieng-Kuntz
	Marc Bourdeau
	Joël Quinqueton

Remerciements

Je tiens tout d'abord à exprimer ma gratitude à Rose Dieng-Kuntz, ma directrice de thèse, pour m'avoir accueilli dans son équipe et avoir assuré le suivi de ma thèse. Ses conseils, qui m'ont été utiles, tant sur l'aspect scientifiques que sur la méthode de recherche, m'ont permis de réaliser ce travail. Particulièrement sa gentillesse et ses encouragements m'ont aidé à surmonter des moments difficiles et de mener à bout cette thèse.

Je tiens à remercier Bruno Fiès et Marc Bourdeau qui m'ont encadré au CSTB, pour leurs nombreux conseils, leur responsabilité et leur soutien pendant tout au long de ces années d'étude.

Je remercie tout particulièrement Madame Parisa Ghodous et Madame Chantal Reynauld d'avoir accepté de rapporter ma thèse. Je les remercie pour pour la rapidité avec laquelle elles ont lu mon manuscrit et pour leurs jugements très pertinents et leurs commentaires sur mon manuscrit, qui ont fait progresser ce document.

Je remercie M. Joël Quinqueton et M. Marc Bourdeau qui m'ont fait l'honneur de participer à mon jury et de s'intéresser à ce travail.

Je veux adresser tous mes remerciements M. Nhan Le-Thanh, qui m'a fait l'honneur de présider mon jury de thèse.

Je tiens à remercier Emmanuelle Loyson pour ce qu'elle m'a apporté lors des séances d'évaluation des algorithmes qui ont été très intéressantes et informatives. Son expérience et ses propositions ont été précieuses pour mes travaux de thèse.

Je suis très reconnaissant envers mes collègues et amis de l'équipe ACACIA : Fabien Gandon, Olivier Corby, Alain Giboin, Sylvain Dehors, Khaled Khelif, Thanh-Le Bach, Laurent Alamarguy et les nouveaux Acaciens. Tout travail scientifique se construit par de petites collaborations. Je garderai un bon souvenir des discussions scientifiques et aussi les atmosphères animées au cours des repas ensoleillés près de la piscine.

La réalisation de ce travail s'appuie également sur un environnement qui est essentiel. A ce titre, je voudrais remercier l'INRIA et son personnel, notamment l'équipe du SEMIR et l'équipe de la documentation. Ces remerciements sont également destinés à Sophie Honnorat et Patricia Maleyran pour leur amitié et leur aide.

Je tiens à remercier le CSTB pour son soutien financier et administratif, durant ces années de recherche.

J'exprime toutes mon amitié au thésards Vietnamiens de l'INRIA, Thanh-Le Bach, Thinh-Van Vu, Lan Le, Trung Nguyen, et beaucoup d'autres qui m'ont beaucoup aidé dans ma vie.

Je termine par un grand remerciement à ma famille, qui est dans mon cœur pour m'avoir encouragé dans les moments difficiles.

Résumé

L'essor d'Internet et du Web a favorisé la mise en ligne de nombreuses informations disponibles, potentiellement utiles pour la veille technologique et scientifique d'une entreprise. Différentes techniques de recherche d'information sur le Web ont été proposées afin de construire des outils permettant d'affiner la recherche pour obtenir des résultats pertinents. Cependant, dans le contexte du Web actuel, malgré de grandes avancées dans le champ de la recherche d'information, ces outils ont montré leurs limites en termes de précision et de rappel.

L'application des technologies du Web Sémantique, en particulier des ontologies, semble donc intéressante pour améliorer les performances de la tâche de veille technologique et scientifique sur le Web. Les travaux de cette thèse se sont déroulés dans le cadre d'une coopération entre le Centre Scientifique et Technique du Bâtiment (CSTB) et l'équipe ACACIA de l'INRIA Sophia Antipolis. L'objectif principal est d'exploiter les technologies du Web Sémantique pour développer un système de veille (OntoWatch), guidé par des ontologies, pour collecter, capturer, filtrer, classer et structurer le contenu du Web en provenance de plusieurs sources d'information dans un scénario d'aide à la veille technologique et scientifique.

Dans une première partie, nous modélisons le processus de veille technologique et scientifique du CSTB reposant sur le modèle général de veille proposé par Lesca. Puis nous identifions les apports potentiels de l'ontologie dans les différentes étapes et nous construisons une ontologie dédiée au système de veille. Cette ontologie intègre une partie d'une ontologie existante et des vocabulaires offerts dans des thésaurus du domaine du CSTB.

Ensuite, nous proposons des algorithmes utilisant une ontologie pour améliorer la recherche des documents sur le Web, puis générer automatiquement les annotations sémantiques (représentees dans le langage RDF) sur ces documents. Ces annotations alimentent dans le système les bases d'annotations, sur lesquelles repose la recherche sémantique d'informations.

Enfin, nous proposons une architecture multi-agents pour l'implémentation du système OntoWatch. Nous nous focalisons en particulier sur la conception des sous-sociétés d'agents dédiées à la recherche et à l'annotation automatique des documents sur le Web.

Mots-Clés : Recherche d'information guidéé par les ontologies, génération d'annotations sémantiques, ontologie, veillle technologique, système multi-agents d'information, RDF(S).

Abstract

The rise of Internet supported the appearance of numerous information available on line, which is potentially useful for the technological and scientific watch of a company. Various techniques of information retrieval on the Web are proposed in order to build tools enabling to refine the search in order to get relevant results. However, in the context of the current Web, in spite of large progresses in the field of information retrieval, these tools showed their limits in terms of precision and recall.

The application of Semantic Web technologies, in particular of ontologies, thus seems to us to be useful to improve the performance of technological and scientific watch task on the Web. This thesis was prepared in the framework of a cooperation between the CSTB (Scientific and Technical Centre for Building) and the ACACIA Team at INRIA Sophia Antipolis. The main objective of this thesis is to use the Semantic Web technologies to develop a system for technology monitoring (OntoWatch). This system is guided by ontologies, in order to collect, capture, filter, classify and structure the Web content coming from several information sources in a scenario of assistance to the technological et scientific watch.

In a first part, we model the CSTB's technological watch process relying on the generic model of monitoring proposed by Lesca. We identify the potential contributions of ontology in the various stages of the process then we build an ontology dedicated to the technological watch system. This ontology integrates a part of an existing ontology and vocabularies offered in thesaurus of the CSTB domain.

After that, we propose several algorithms using an ontology to improve document search on the Web and to generate automatically semantic annotations (in RDF format) for these documents. These annotations feed the annotation bases of the system, bases on which the semantic search of information relies.

Finally, we propose a multi agents architecture for implementation of the OntoWatch system. We focus in particular on the design of the sub-societies of agents dedicated to search and automatic annotation of documents on the Web.

Keywords : Semantic information retrieval, semantic annotations generation, ontology, technology monitoring, technological watch, information multi-agents system, RDF(S).

Table des matières

Liste des figures

Introduction

Contexte scientifique et industriel

Depuis plus d'une dizaine d'années, la veille technologique est de plus en plus intégrée dans les entreprises. Face à la transformation rapide et profonde du monde de la science et de la technologie, afin de protéger son avenir, l'entreprise doit surveiller son environnement pour constamment prendre connaissance de tout ce qui évolue autour d'elle. En possédant des informations clés, l'entreprise peut alors anticiper les changements et prendre de bonnes décisions. Le besoin incessant d'être informé des dernières inventions, innovations, et des nouvelles technologies a conduit à la mise en place de services de veille ainsi qu'à plusieurs recherches sur les méthodes de collecte et de traitement de l'information en provenance de l'extérieur.

Le Web actuel connaît un succès impressionnant avec l'énorme quantité d'information disponible et en croissance exponentielle. Avec l´essor du commerce électronique et des publications institutionnelles en ligne, de nouveaux types d'informations deviennent accessibles pour tous, comme par exemple : la description d´une société, sa structure, ses produits, etc, mais aussi des précisions

sur les normes et le cadre législatif d´un pays, des rapports annuels, des informations financières ou encore les thèmes de recherche d´un laboratoire. Le Web est actuellement de manière indéniable la plus grande source d'information électronique et constitue une formidable mine d'or pour les activités de veille technologique. Construire des système de veille afin d'exploiter efficacement ces ressources est donc de plus en plus nécessaire, pour les entreprises et les organisations.

Différentes techniques de récupération d'information sur le Web ont été proposées afin de construire des outils permettant d'affiner la recherche pour dégager des résultats pertinents. Cependant, comme le Web actuel reste encore syntaxique et lisible mais non compréhensible par les machines, malgré de grandes avancées dans le champ de la recherche d'information, ces outils ont montré leurs limitations en termes de précision et de rappel.

L'apparition d'une nouvelle génération du Web, nommé « Web Sémantique », est un des efforts pour pallier ces limitations. Reposant sur l'idée d'utiliser les langages de représentation des connaissances pour modéliser le contenu sémantique des ressources du Web, le Web sémantique promet de rendre le Web compréhensible à des machines. Cette approche s'appuie sur l'explicitation de la conceptualisation d'un domaine, partagée par une communauté et représentée dans une ontologie du domaine concerné. Avec l'aide de l'ontologie, les moteurs de recherche sémantique peuvent faire des inférences et des raisonnements sur les annotations sémantiques pour obtenir des résultats plus pertinents. Ces annotations sémantiques sont des métadonnées décrivant des ressources du Web en utilisant des vocabulaires appartenant à ces ontologies. Néanmoins, il reste plusieurs problèmes de recherche pour réaliser cette vision et les applications du Web Sémantique sont encore dans la période initiale du développement.

L'application des technologies du Web Sémantique, en particulier des ontologies, semble donc une clé pour améliorer les performances de la tâche de veille technologique et scientifique sur le Web. C'est dans ce contexte que s'est formée

une coopération entre le Centre Scientifique et Technique du Bâtiment (CSTB) et l'équipe ACACIA de l'Inria Sophia Antipolis. Les travaux de cette thèse se sont déroulés dans le cadre de ce partenariat, dont un des principaux objectifs est d'exploiter les technologies du Web Sémantique pour développer un système multi-agents de veille, ces agents étant guidés par des ontologies, pour collecter, capturer, filtrer, classer et structurer le contenu du Web en provenance de plusieurs sources d'information dans un scénario d'aide à la veille technologique et scientifique.

Problèmes posés et objectifs poursuivis

Comment trouver de l'information ? Comment gérer des informations trouvées d'une manière plus efficace pour des exploitations ultérieures ? Comment ne pas perdre des informations pertinentes ? Notre problématique de recherche dans le domaine de la veille démarre avec ces questions.

Dans le cadre de cette thèse, nous approfondissons les problèmes suivants :

- La détection de bonnes sources d'informations sur le Web (externe et interne) : Il s'agit de déterminer différentes sources d'informations à interroger. Comment exploiter les caractéristiques de ces sources pour aider le processus de recherche d'information ?
- L'hétérogénéité de sources : selon le cas, elles pourront être structurées, semi-structurées, ou non structurées.
- L'application de la recherche d'information sémantique qui repose sur des annotations sémantiques dans le système de veille. Ce problème se traduit en plusieurs questions à résoudre : comment disposer d'annotations de ressources sur le Web, comment travailler avec des sources non annotées et construire automatiquement leurs annotations si leurs auteurs ne les ont pas annotées ? Et comment construire une ontologie pouvant servir de ressource conceptuelle dédiée à la veille technologique ?
- L'amélioration de la recherche d'information classique grâce à une ontologie la guidant pour réduire le manque des informations pertinentes dans les

résultats.

Nous reposerons sur les hypothèses de base suivantes :

- Les technologies du Web Sémantique sont utiles pour construire un système de veille facilitant les tâches de recherche et de traitement d'informations du veilleur.
- L'utilisation de l'ontologie peut être utile dans plusieurs étapes du processus de veille et permet d'obtenir de meilleurs résultats par rapport aux approches courantes qui reposent sur l'utilisation des outils de recherche d'information.

En réponse au besoin exprimé par les veilleurs du CSTB et aux problèmes posés par la veille technologique et scientifique sur le Web, nous nous sommes fixé les objectifs suivants :

- Exploiter les techniques du Web Sémantique dans les étapes du processus de veille au CSTB où ces techniques semblent applicables.
- Proposer une architecture pour construire un système d'information d'aide à la veille technologique.

Contributions et champs de recherche concernés

Comme réponse à cette problématique, nous proposons le système d'aide à la veille appelé OntoWatch qui repose largement sur une ontologie et sur les techniques du Web Sémantique. Ce système permet aux veilleurs d'exprimer leurs besoins d'informations, puis de collecter des informations pertinentes sur le Web externe et interne, ensuite les intégrer dans le système sous forme des annotations sémantiques. Il fournit un accès aux résultats de veille via les mécanismes de recherche d'information sémantique. Concrètement, la thèse :

- Modélise le processus de veille technologique et scientifique du CSTB reposant sur le modèle général de veille proposé par Lesca. A partir de cela, nous avons identifié les apports potentiels de l'ontologie dans les diverses

étapes.

- Introduit des expériences dans la tâche de construction d'une ontologie qui va guider le système de veille. Cette ontologie permet non seulement de modéliser les connaissances dans les domaines de la veille, mais aussi les concepts importants dans la tâche de veille.
- Propose une architecture du système de veille, reposant sur l'approche Web Sémantique et sur une ontologie, et sur le paradigme de plusieurs agents coopérant.
- Propose des algorithmes pour la recherche d'information sur le Web et la génération automatique des annotations sémantiques en utilisant l'ontologie.

Le système OntoWatch vise à faciliter les travaux du veilleur dans le domaine de la construction et du bâtiment mais il devrait pouvoir être facilement adapté à d'autres domaines de veille, puisque l'approche et les algorithmes proposés sont complètement génériques.

Nos travaux concernent trois domaines de recherche principaux :

Les technologies du Web Sémantique : Nous avons adapté l'approche Web Sémantique d'entreprise de l'équipe Acacia au scénario de la veille technologique et scientifique. Les connaissances du domaine de veille, les résultats du système de veille sont matérialisés en utilisant l'ontologie et les annotations sémantiques. La recherche d'information repose ensuite soit sur la recherche sémantique d'information, soit sur la recherche étendue guidée par l'ontologie.

Les technologies multi-agents : Avec l'autonomie de chaque individu et la distribution des travaux entre eux dans une société, le paradigme multi-agents semble approprié pour l'implémentation des composants d'un système de veille, qui travaille avec des ressources hétérogènes et distribuées.

Organisation du document

Ce mémoire de thèse se compose de neuf chapitres répartis en deux parties. La première partie est l'état de l'art des domaines de recherche concernant la veille, le

Web Sémantique et les systèmes multi-agents. La deuxième partie de thèse est dédiée à la description du processus de construction d'OntoWatch, notre système d'aide à la veille technologique et scientifique guidé par une ontologie.

Le chapitre 1 présente un panorama général de la veille, notre cadre privilégié d'application. Nous y abordons les concepts de base, notamment le concept de la veille technologique sur le Web. Ensuite, nous nous focalisons sur les sources et les types d'information - la matière première du processus de veille.

Le chapitre 2 introduit la vision du Web Sémantique, l'évolution du Web actuel, en focalisant sur ses composants principaux. Puis nous nous intéressons à l'aspect recherche d'information reposant sur les techniques du Web Sémantique et nous analysons plusieurs méthodes et outils d'extraction d'information pour l'annotation automatique. Enfin, nous abordons les systèmes multi-agents et leurs applications dans le contexte de la recherche d'information.

Le chapitre 3 présente la situation actuelle de la veille technologique et scientifique au CSTB. Tout d'abord nous décrivons les caractéristiques du système de veille courant et les problèmes sous-jacents. Ensuite, en nous basant sur le modèle de Lesca, nous modélisons le processus de veille au CSTB, permettant de chercher à quel point les améliorations peuvent être réalisées.

Le chapitre 4 décrit les travaux réalisés pour la construction de l'ontologie O'Watch, le coeur du système de veille. Nous expliquons comment réutiliser une ontologie existante et comment transformer des vocabulaires offerts dans des thésaurus pour la mise en oeuvre d'une ontologie dédiée à la veille technologique, formalisée en RDF(S).

Le chapitre 5 présente notre approche générale pour le développement du système OntoWatch, l'architecture logique de ce système.

Dans le chapitre 6, nous présentons une partie importante de nos travaux, les algorithmes utilisant une ontologie pour améliorer la recherche des documents sur le Web, puis générer automatiquement les annotations sémantiques de ces documents. Ces annotations alimentent les bases d'annotation dans le système, sur lesquelles

repose la recherche sémantique d'information.

Le chapitre 7 propose une architecture multi-agents pour l'implémentation du système OntoWatch. Nous nous focalisons en particulier sur la conception des sous-sociétés d'agents dédiées à la recherche et à l'annotation automatique des documents sur le Web. Nous expliquons les étapes principales de la conception incluant l'analyse du fonctionnement, la description de rôles d'agents et leurs protocoles d'interaction.

Le chapitre 8 présente l'évaluation de la recherche d'information dans le système proposé.

Pour conclure ces travaux de thèse, nous résumons dans le chapitre conclusions et perspectives, nos principales propositions et les problèmes que nous avons résolus avant de discuter les points restant à approfondir dans le futur.

Partie I:

Etat de l’art

1 La veille sur le Web

Dans un monde de la globalisation de l'économie, des échanges économiques autant que la transformation rapide et profonde des sciences et techniques, l'entreprise moderne, pour s'adapter, rester compétitive et prospérer, doit anticiper les évolutions de son environnement. L'information est au cœur d'une telle démarche d'intelligence stratégique, technique, économique et sociale. L'entreprise doit être constamment informée des dernières découvertes, inventions ou innovations. Elle doit pour cela s'imposer une constante observation des mutations scientifiques, techniques et technologiques.

Dans les dernières années, l'essor d'Internet a favorisé l'apparition de nombreuses informations disponibles en ligne et le Web promet d'être une mine d'or pour toutes les organisations. En effet, un grand nombre de documents publics et disponibles sur Internet (dépêches de presse, bases de données bibliographiques scientifiques et techniques, etc.) ou en Intranet (mails électroniques, rapports techniques, rapports d'activité) contiennent potentiellement de l'information utile à la décision. Le volume croissant et l'hétérogénéité des informations posent de vrais obstacles à dépasser pour efficacement exploiter cette mine d'or.

Dans ce contexte, la veille sur le Web est aujourd'hui un processus vital pour une entreprise. Le besoin d'être correctement informé et alerté est de plus en plus évident, et donne naissance à un domaine professionnel : la veille, avec en particulier la veille stratégique, la veille technologique, la veille concurrentielle.

1.1 Qu'est ce que la veille ?

Le terme « veille » est souvent utilisé pour désigner cette activité de surveillance de l'environnement des entreprises. Sous un effet de mode, une récente prolifération d'adjectifs est venue qualifier le terme de veille. On trouve ainsi les appellations veille industrielle, veille globale, veille environnementale, veille stratégique, veille informative, veille technologique, veille concurrentielle, veille commerciale, veille d'acquisition, veille des ressources humaines... Cette terminologie est employée selon des pratiques assez confuses. Toutes ces veilles ne sont pas censées couvrir les mêmes activités.

On peut définir la veille générale comme l'ensemble des activités de surveillance sur l'environnement d'une entreprise pour fournir des données utiles à la définition de ses stratégies d'évolution. Cette surveillance récolte donc des informations de natures très variées: économique, financière, commerciale, scientifique, technique, technologique, sociologique, politique, juridique, les clients, les sous-traitants, les fournisseurs... Nous avons préféré l'appellation veille industrielle (utilisée par Martinet et Ribault dans [Martinet et Ribault, 1989]) à ses synonymes veille globale, veille environnementale, veille informative, veille stratégique parce qu'elle paraît mieux retranscrire la réunion de l'ensemble de ces activités et surtout parce qu'elle précise à qui elle s'adresse.

Nous faisons le point sur les différents aspects de la veille et présentons en détail un aspect important concernant la thèse : la veille technologique.

1.1.1 Typologie de veille

Il existe plusieurs critères pour différencier les types de veille. Nous présentons ici la typologie de veille selon le domaine d'application, l'horizon de temps sur lequel porte la recherche d'informations, le mode de conduite de la recherche et l'attitude culturelle face à l'information.

1.1.1.1 Domaines d'application

Les divers types de veille sont souvent regroupés en quatre grands domaines d'application: veille technologique, veille concurrentielle, marketing ou commerciale, veille sur les autres facteurs de l'environnement (veille sociale, environnementale, …).

La veille technologique

Cette veille consiste à surveiller : les dépôts de brevets, l'évolution de normes, l'évolution des technologies, les ruptures technologiques, les procédés de fabrication, la recherche fondamentale, les articles scientifiques, les thèses, les rapports scientifiques.

Cette activité est généralement située dans la direction Recherche et Développement. Son objectif est d'aider à améliorer les produits existants, créer de nouveaux produits, éviter d'être contrefacteur, connaître les projets des concurrents. Ses clients sont les ingénieurs des bureaux d'études et de production. La collecte de l'information est souvent du ressort du centre de documentation aidé par les ingénieurs de la propriété industrielle pour les brevets.

La veille concurrentielle

Cette veille permet de connaître parfaitement ses concurrents directs. Il s'agit de disposer d'une description la plus complète possible : identification (adresse, forme juridique), activité, effectifs, marques gérées, résultats financiers, rapport d'activité, publicité, productions, investissements, projets. Une bonne partie des informations est directement collectée par la direction Marketing en association avec la direction Financière. Le centre de documentation est cependant sollicité pour compléter par des sources formelles les informations sur les marques, la production et les projets.

La veille commerciale

Elle consiste à collecter des informations sur ses propres clients, leurs taux de satisfaction, leurs besoins, connaître la part de marché du produit, connaître les marges des produits concurrents ainsi que leurs techniques de vente. Réalisée

généralement au sein de la direction Commerciale, la veille commerciale est un produit de l'activité du service Marketing et Commercial.

La veille sociale

Elle permet de suivre la réglementation sociale en cours et à venir par la présence dans des groupements professionnels. La direction de ressources humaines prend cette activité à son compte car elle a une grande pratique de l'analyse des textes juridiques.

La veille environnementale

Elle concerne la protection de l'environnement. Cette contrainte est actuellement forte. Elle fait partie des soucis de la recherche et du développement car ce problème concerne non seulement la production mais aussi la destruction du produit. Cette veille peut être associée à la veille technologique.

La veille juridique

Elle concerne les législations et règlements nationaux, européens et internationaux pour avoir une incidence sur l'offre ou sur les modes de fonctionnement de l'entreprise.

La veille géopolitique

Elle s'intéresse à l'environnement international, aux risques politiques, sociaux et économiques des pays instables ou des marchés émergents.

La veille stratégique

Elle coordonne les différentes veilles. Elle peut présenter une ou plusieurs facettes appelées : veille technologique, veille commerciale, veille concurrentielle, etc. Compte tenu de la nature des informations concernées, elle s'apparente au traitement du signal. Elle participe à l'information de la direction pour ses choix de développement. Elle exige une organisation globale, des outils appropriés et un animateur coordinateur. La veille stratégique est le processus informationnel par lequel l'entreprise peut s'informer de l'état et de l'évolution de son environnement économique en vue de survivre avec succès.

[Lesca et Blanco, 2002] a donné la définition suivante de la Veille Stratégique : « La Veille Stratégique est le processus par lequel un individu ou un groupe d'individus traquent, de façon volontariste, et utilisent des informations à caractère anticipatif concernant les changements susceptibles de se produire dans l'environnement extérieur dans le but de créer des opportunités d'affaires et de réduire des risques et l'incertitude en général ».

Finalement l'objectif de la veille stratégique est de permettre d'agir très vite et au bon moment. La capacité de survie des organisations dépend pour partie de leur aptitude à anticiper les changements extérieurs et à s'adapter.

1.1.1.2 Horizon de temps sur lequel porte la recherche, le mode de la recherche et l'attitude face à l'information

La séparation des diverses veilles basées sur la nature des informations traitées, favorise une exploitation localisée des données au détriment d'une vision globale : par exemple, une veille technologique nécessite la connaissance des acteurs du marché, notamment ceux dont la technologie domine. Il s'ensuit des problèmes de circulation de l'information, de redondance, de perte d'efficacité. L'hétérogénéité de la population concernée et la diversité des problématiques expliquent la complexité de la mise en oeuvre d'une approche globale de la veille.

[Degoul, 2001] propose d'adopter un angle d'observation nouveau basé sur la performance globale. Il privilégie trois axes d'analyse : l'horizon de temps sur lequel portent la recherche d'informations, le mode de conduite de la recherche et l'attitude culturelle face à l'information.

Trois horizons de temps sur lequel porte la recherche caractérisent les informations recherchées :

- Recherche d'informations rétrospectives : la recherche concerne des évènements totalement réalisés, des faits accomplis.
- Recherche d'informations concernant le passé récent : La recherche concerne des évènements qui viennent juste de se produire ou qui sont en cours de réalisation.
- Recherche d'informations concernant le futur : L'objectif est de repérer,

cette fois sur un horizon de temps probabilisable (au-delà nous rentrerions dans la prospective), des évènements en préparation, non encore réalisés.

Les modes de conduite de la démarche de la recherche d'information sont :

- Recherche ponctuelle : La recherche est menée pour répondre à une préoccupation immédiate. L'objectif étant de trouver une réponse, elle s'arrête lorsque les éléments de réponse sont trouvés.
- Recherche permanente : La recherche est menée de façon récurrente. L'objectif est ici de suivre au « fil de l'eau » les évolutions des composantes de l'environnement.
- Recherche en vue de formuler des conjectures : La finalité de certaines recherches est de conduire à des suppositions fondées sur des indications, des apparences, des probabilités d'occurrence d'évènements. Il s'agit d'exprimer des hypothèses qui n'ont encore reçu aucune confirmation. L'intérêt de ces hypothèses s'estompe avec le temps : il s'arrête lorsque l'hypothèse est confirmée ou infirmée.

L'attitude culturelle face à l'information est distinguée en deux niveaux suivants :

- Individualisation : Très souvent, l'information est captée d'une façon très individualiste et reste peu diffusée.
- Mutualisation : Dans d'autres cas, plus rares, l'information collectée à l'extérieur est portée à la connaissance du plus grand nombre d'acteurs concernés.

En se basant sur ces éléments d'analyse, [Degoul, 2001] définit les trois approches opérationnelles de la veille:

La veille spontanée

Elle répond à un besoin clairement perçu, explicitement formulé. Ce type de veille est guidé par la recherche ponctuelle d'informations ou par une recherche sur une thématique bien ciblée. Il s'agit ici de répondre à des questions du type « Qui fait quoi? Comment faire? Qui sait ou fait quoi ? », d'identifier, d'établir des repères ou

de comprendre l'état actuel de l'environnement. C'est une mission de « prestation à la demande » ou de « questions/réponses» en vue de connaître l'environnement ou de résoudre des problèmes. La veille spontanée s'appuie sur un « état » (aspect statique) concernant les acteurs de l'environnement. Elle est sollicitée par toutes les fonctions et tous les niveaux hiérarchiques de l'entreprise.

La veille réactive

Elle est caractérisée par une surveillance permanente sur un thème clé. Sa mise en place se fait en réaction à un échec majeur qui a affecté la rentabilité ou parfois la survie de l'entreprise, ou à un environnement qui évolue trop rapidement, et ceci pour pouvoir infléchir, adapter, améliorer les politiques et la stratégie de l'entreprise en temps voulu. Elle vise à modifier le système de l'entreprise, contribuer à transformer les règles qui le régissent, modifiant les hypothèses habituellement admises dont elles sont issues et contribue à la transformation des mentalités et des comportements. Elle est essentiellement centrée sur le passé récent : événements en cours ou qui viennent juste de se réaliser. Il s'agit donc essentiellement d'une information « de suivi », c'est-à-dire d'une information émergente. C'est de la permanence de la démarche que dépend la possibilité de détecter l'information au plus tôt, de la saisir le plus en amont possible dès l'apparition de l'événement. La veille réactive s'intéresse aux évolutions, aux changements (aspect dynamique).

La veille anticipative

Elle est liée à la volonté de rechercher une vision du futur proche. Il s'agit d'anticiper les événements majeurs, les changements rapides, les ruptures de tendance « non prévisibles » à partir des continuités non prévues (les ruptures). Une démarche spécifique, basée sur une information d'alerte constituée de signaux précurseurs, annonciateurs de changement (signaux faibles) est alors nécessaire.

1.2 Approche théorique de la veille technologique

Nous présentons ici ce qu'est la veille technologique. Avant d'exposer la nature de l'activité de la veille technologique, nous allons donner une définition de ce que

l'on entend par technologie. Reprenons simplement l'énoncé qu'en a fait F Lainé dans [Laine, 1991]:

"La technologie, c'est l'ensemble des connaissances scientifiques et des savoir-faire applicables aux arts industriels."

On peut la différencier de la technique par le simple fait qu'elle rentre dans un processus industriel, c'est-à-dire dans un processus de transformation d'un produit de façon à lui donner de la valeur ajoutée.

1.2.1 Définition de la veille technologique

Il existe plusieurs définitions de la veille technologique. En s'appuyant sur les points de vue d'experts en veille technologique, nous en proposons ici quelques-unes, dans un ordre chronologique.

Voici une première définition présentée par Henri Dou et François Jakobiak en 1995 [Dou et Jakobiak, 1995] :

La veille technologique (ou veille) est « l'observation et l'analyse de l'environnement scientifique, technique, technologique suivie de la diffusion bien ciblée, aux responsables, des informations sélectionnées et traitées, utiles à la prise de décision stratégique ».

Daniel Rouach présente plusieurs définitions de divers auteurs, en 1996 [Rouach, 1996,] :
Pour Steven C. Wheelwright, elle est *« constituée par l'ensemble des techniques visant à organiser de façon systématique, la collecte, l'analyse, la diffusion de l'exploitation des informations techniques utiles à la sauvegarde et à la croissance des entreprises ».*
Pour R. Beaussier, de la société CEGELEC : la veille technologique est *« l'exploitation systématique et surtout organisée de l'information industrielle. Cette technique de veille technologique consiste à savoir écouter et regarder pour repérer toutes les innovations utiles assurant l'aide aux développements techniques indispensables à l'entreprise face à la concurrence mondiale ».*
Dans la revue La Recherche : *« La veille technologique est le moyen pour l'entreprise de faire émerger les éléments stratégiques de la masse*

d'informations disponible aujourd'hui. Ni espionnage industriel, ni réalisation d'un état de l'art purement spéculatif dans un domaine technique restreint, la veille est avant tout destinée à éclairer les responsables de l'entreprise dans la résolution des problèmes industriels auxquels ils sont confrontés. »

Enfin, Henry Samier et Victor Sandoval [Samier, Sandoval, 1998, p. 115] proposent cette autre définition :

« La veille technologique est un ensemble de techniques licites visant à organiser de façon systématique la collecte, l'analyse, la diffusion et l'exploitation des informations technologiques utiles à la sauvegarde et à la croissance des entreprises par la conception et le développement de produits. »

D'un point de vue synthétique, la veille technologique est un ensemble de techniques ou une activité licite(s) qui consiste (nt) à observer (surveiller, écouter, regarder) l'environnement scientifique, technique et technologique, afin de repérer, collecter, organiser, analyser, exploiter puis diffuser les informations utiles qui vont permettre : d'anticiper les évolutions ; la sauvegarde et la croissance ; la résolution de problèmes industriels ; d'aider aux développements techniques.

1.2.2 Acteur de la veille technologique

Alors que l'expression «veille technologique» fait l'objet de nombreuses définitions, l'acteur principal dans la veille, appelé le veilleur, n'est jamais décrit précisément. En fait, ce n'est généralement pas une mais plusieurs personnes qui doivent intervenir dans une démarche de veille. Par contre, plusieurs profils sont présentés.

Par exemple, selon Jakobiak [Jakobiak, 1995], trois réseaux de spécialistes interviennent dans une démarche de veille :

- Dans un premier temps, un réseau d'observateurs est chargé de la recherche, de la collecte et de la diffusion de l'information brute sélectionnée ;
- Dans un deuxième temps, un réseau d'experts analyse les informations

obtenues. Ces experts peuvent être répartis dans quatre familles : application, produits, procédés, prospective.

- Enfin, un réseau de décideurs intervient dans un troisième temps.

Donc, la veille n'est pas la tâche d'une personne, mais celle d'un ensemble d'individus. Cet ensemble va être plus ou moins complet selon les moyens et la taille de l'entreprise. Il est même possible de considérer chaque personne de l'entreprise comme étant un veilleur, car chacun peut capter des informations dans des revues, sur des salons, etc. et les transmettre aux décideurs.

Trois principaux groupes d'acteurs apparaissent donc dans ce schéma :

Les décideurs pilotent le dispositif de veille. Leur rôle se situe en aval et en amont du processus de veille. Ils définissent les axes prioritaires de la veille et utilisent les remontées d'information élaborée, afin de prendre des décisions d'ordre stratégique.

Les veilleurs, spécialistes, font remonter à leurs clients (techniciens, ingénieurs, chercheurs, décideurs,…) des informations à valeur ajoutée. En cherchant à regrouper les caractéristiques importantes d'un veilleur, [Goujon, 2000] donne pour caractéristiques importantes d'un veilleur:

- savoir-faire méthodologique, organisation du travail en équipe ;
- connaissances des sources d'information et des outils de recherche d'information (Internet, banques de données, revues, ...) ;
- expertise scientifique et technique sur le sujet ou le domaine de veille ;
- connaissances générales sur le marché, la concurrence, les innovations, les travaux en cours, ... ;
- connaissances sur l'entreprise : son évolution, ses capacités d'innovation, ses points faibles,... ;
- connaissances informatiques, statistiques : utilisation d'outils bibliométriques, interprétation des résultats.

Les animateurs intermédiaires entre les spécialistes et les décideurs. Le rôle d'animateur est à facettes multiples :

- Organiser l'activité : méthode de travail, réunion, coordination des analystes,

- Fédérer le travail effectué dans les différents groupes.
- Mesurer l'activité et évaluer les besoins,.

Les experts (ou analyseurs) interviennent lors de certaines étapes du processus de veille. Ils ont une parfaite connaissance de leur domaine qui leur permet de faire des recommandations en s'appuyant sur l'analyse et la validation des informations. Ils peuvent également être diffuseurs d'une information d'alerte qu'ils sont les seuls à connaître. Ce sont eux qui valident la pertinence de l'information collectée et font des comptes-rendus aux décideurs (fiches synthèses, propositions d'actions)

Un veilleur est dans plusieurs cas un expert du domaine de veille, car c'est lui qui analyse les documents ou informations techniques et qui en extrait des connaissances stratégiques. Dans notre démarche, nous avons considéré ces veilleurs-experts comme les utilisateurs du système d'aide à la veille que nous devons réaliser.

1.3 Source d'information nécessaire

Dans une démarche de veille technologique, les sources d'information exploitées sont importantes, puisque les résultats qui vont être obtenus dépendent fortement de leur qualité générale ou de leur pertinence spécifique vis-à-vis du sujet.

Afin de mieux cerner le concept d'« information », nous présentons dans un premier temps plusieurs typologies de l'information, qui sont présentées par des spécialistes de la documentation et de la veille : J. Chaumier, F. Jakobiak et D. Rouach, Humbert Lesca,... L'information est ainsi caractérisée en fonction de son niveau d'élaboration (ou classe ou catégorie d'information spécialisée), du media, du type d'information (information courante ou stratégique), de sa nature (scientifique ou économique), du support (papier ou magnétique), etc.

Parmi les sources d'information électroniques, les bases de données payantes en ligne sont des sources fiables mais coûteuses, tandis qu'Internet représente une source peu coûteuse mais pas toujours fiable.

Enfin, parmi les documents utilisés pour faire de la veille technologique, le brevet constitue le support d'information le plus exploité pour la veille technologique. Les brevets décrivent en effet toutes les innovations, et représentent

ainsi la principale source d'information technique et stratégique : si le nombre des brevets déposés dans un secteur donné est élevé, c'est qu'une action importante est en cours.

1.3.1 Typologie de l'information

La définition d'une typologie de l'information est indispensable pour tout activité de collecte et de traitement de l'information. L'information est typée en fonction des plusieurs critères, et des points de vue différents. Nous en présentons ici quelques unes.

Information blanche, grise, noire

Selon leur degré d'accessibilité, selon les codes couleurs adoptés par la norme XP X50-053, l'AFNOR distingue trois types d'information : blanche, grise, noire

- « L'information aisément et licitement accessible » que certains appellent « l'information blanche ». Ouverte à tous, son accès ne présente aucun problème juridique. Elle se trouve dans la presse, les publications spécialisées, les colloques, la littérature grise, les banques de données, Internet…

- « L'information licitement accessible mais caractérisée par des difficultés dans la connaissance de son existence et de son accès ». Cette « information grise » se présente sous une forme plus élaborée. Pour la trouver, il faut d'abord savoir la chercher. Elle se rapproche davantage du renseignement.

- « L'information à diffusion restreinte et dont l'accès et l'usage sont expressément protégés ». Il s'agit ici de « l'information noire » qui est protégée par un contrat ou une loi. Seules quelques personnes sont autorisées à y accéder. Son recueil peut entraîner des poursuites pénales si une personne non autorisée la détient.

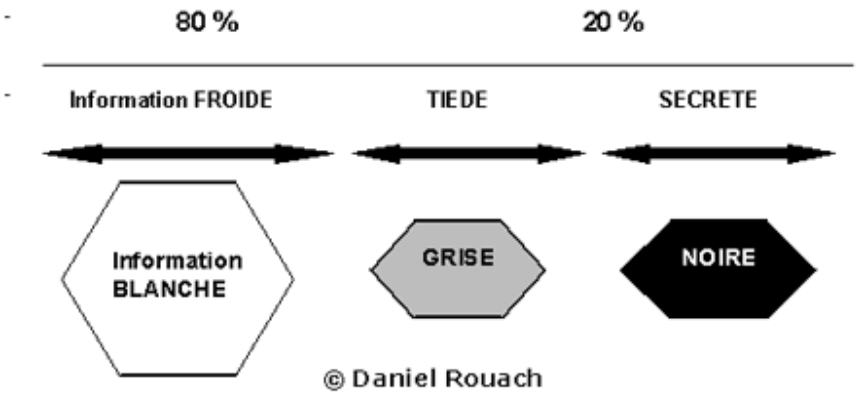


Figure 1 Information blanche, grise, et noire

Information formelle, informelle

C'est selon le support physique de l'information que [Martinet et Ribault, 1989] distinguent l'information formelle de l'informelle qui correspond à l'information orale.

L'information formelle est écrite pour être lue (publiée sur un support papier ou électronique et directement exploitable). Issue de publications, normes, brevets, lois, documents divers de l'entreprise, presse, bases de données, internet, l'information formelle (information de type « texte ») est répertoriée, indexée, codifiée. On estime que l'information technique est retrouvée à hauteur de 80 % dans les brevets. L'information scientifique disponible presque intégralement sous forme formelle.

L'information informelle est non transcrite et dissociée de tout support tangible (généralement générée par l'entreprise et ses acteurs ou obtenue via les clients, partenaires, concurrents, experts, et réseaux de contacts personnels). Sans support précis, souvent orale et qualitative, l'information informelle nécessite un grand travail d'analyse, de vérification et de recoupement avec d'autres sources pour être retenue et utilisée par l'entreprise. L'information commerciale est majoritairement présente sous cette forme, Il s'agit d'une information souvent générée par l'entreprise elle-même, pas ses contacts avec les clients, les fournisseurs, les experts, les concurrents, les partenaires politiques. Ce type d'information peut être relevé également sur les forums ou sur les listes de diffusion.

L'information informelle comprend l'information de type « floue », l'information de type « expertise » et l'information de type « foires et salons ». La même distinction s'applique aux sources d'information, formelles (presse généraliste et spécialisée, revues scientifiques, ouvrages, banques de données, brevets, Internet) ou informelles (clients, concurrents et partenaires ; missions et voyages d'études ; expositions et salons ; colloques, congrès et clubs ; étudiants, stagiaires et thésards ; réseau de contact personnel).

Information ouverte / fermée

Pour [Baumard, 1991] si « la source délivre l'information de son plein gré, cette information est ouverte, sinon elle est fermée ». Pour d'autres spécialistes, « l'information fermée correspond à ce qui n'est pas publié. L'information ouverte est assimilée à l'information écrite». L'information ouverte représente donc une information de « premier niveau » par opposition à une information à haute valeur ajoutée et confidentielle. Très abondante, l'information dite "ouverte", pour être efficace et utile pour la prise de décision, doit être répertoriée, décomposée, retravaillée, complétée, classée et diffusée. Quant à l'information fermée, elle émane des circuits propres à l'entreprise. Elle peut parfois être obtenue de manière illégale. La collecte de cette information non formalisée nécessite la mise en place d'une méthode.

Information primaire / secondaire

L'information primaire est issue des documents primaires : livres, textes de loi, brevets, articles … L'information secondaire recense et/ou analyse une information primaire. Il s'agit d'une notice bibliographique enrichie ou non d'un résumé. Les bibliographies et catalogues constituent des documents secondaires.

Information brute / élaborée

L'information brute est telle qu'elle a été obtenue au sens où elle n'est ni qualifiée ni interprétée. Il s'agit d'information sans valeur ajoutée (même si elle n'est pas sans valeur). Les informations brutes sont comparables à des données car elle ne

sont pas contextualisées et leur signification n'est pas prise en compte.

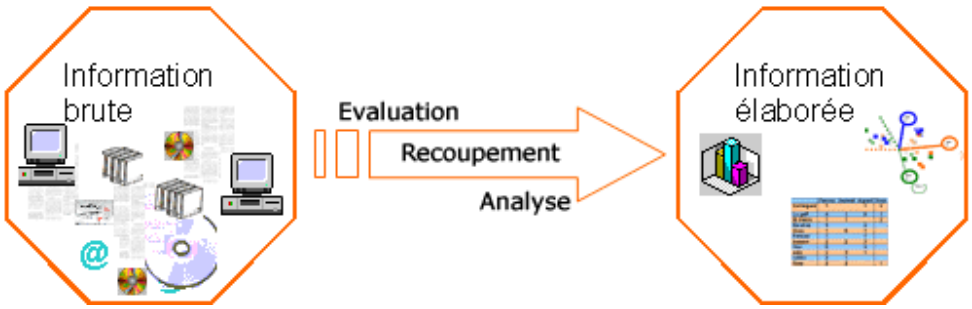


Figure 2 Information brute, élaborée

L'information élaborée est à valeur ajoutée car contextualisée, qualifiée, recoupée et validée.

Information explicite / tacite

L'information explicite revêt la forme d'un document formalisé : une méthode, un processus, un fichier clients ou fournisseurs, un mode d'emploi. Facilement transmissible, cette information peut apparaître et se diffuser dans le système d'information de l'entreprise.

L'information tacite est une information non formalisée puisqu'elle concerne les tours de main, les « trucs », les secrets de métier qu'un salarié a acquis et échangé lors de ses contacts à l'intérieur et l'extérieur de son organisation. Elle est perçue par la pratique et dans l'action. Plus intuitive, l'information tacite est difficilement identifiée et donc transmissible.

De plus, [Lesca et Elisabeth, 1999] a dressé trois types d'information de l'entreprise.

L'information de fonctionnement

Il s'agit des informations indispensables ou presque au bon fonctionnement « mécanique » quotidien de l'entreprise. Ce sont souvent des tâches répétitives. Elles concernent les opérations de commande et de contrôle.

Commande : Les informations de commande sont nécessaires à la réalisation d'une commande ou d'une tâche bien définie : la commande du client, celle du fournisseur, la fiche produit… Cette opération, souvent répétitive, est généralement la première à être informatisée dans une entreprise.

Contrôle : Les opérations de contrôle se révèlent être davantage le fruit d'une volonté au sein de l'entreprise. Les informations de contrôle interviennent pour contrôler les résultats d'une opération, d'une tâche, d'un évènement : fiche de stock, bilan de l'entreprise, relevé bancaire, historique des ventes…

L'information d'influence

Ici « influence » se rapproche des termes : animation, stimulation, motivation, coordination. L'information d'influence est le ciment de la cohésion de l'entreprise. Sa finalité est d'influer sur le comportement des acteurs de l'entreprise en interne mais aussi à l'extérieur. Elle prend la forme pour les membres de l'entreprise de bruits de couloir, journal interne, réunions… et pour l'extérieur se présente sous forme de plaquettes publicitaires, catalogue produits, rapports adressés à la banque… Elle peut donc se présenter de manière totalement informelle ou au contraire très formalisée.

L'information d'anticipation

Visée par la veille, ce type d'information regroupe les informations qui alimentent le pilotage et permet ainsi à l'entreprise « de voir venir à l'avance certains changements de son environnement socio-économique dans le but d'en tirer un avantage ou bien d'éviter un risque ». Il s'agit d'informations concernant un concurrent, d'un renseignement concernant le projet d'un client, d'une rumeur concernant une modification dans la législation, la parution d'un article scientifique…

Selon Charles Zartarian et Vahé Hunt [Hunt et Zartarian, 1990] l'information peut être divisée en quatre grands ensembles importants :

L'information de type « texte »

Il s'agit de l'information structurée provenant généralement des bases de données internes ou externes (serveurs, internet) : normes, contraintes environnementales, brevets. Cette information est validée, travaillée, codifiée afin de la rendre accessible.

L'information de type « floue »

Ce type d'information concerne en priorité des informations glanées à l'extérieur de l'entreprise par le personnel en contact avec les clients, les fournisseurs, les commerciaux après un salon, les experts après un colloque : l'existence d'un réseau de personnes débouche sur l'apparition des ces informations. Elles sont très souvent éparpillées dans les différents services de l'entreprise : achat, marketing, R&D, vente. Pour identifier ce type d'information, généralement sans support formel, il est primordial de mettre en place une stratégie de collecte, un processus de contrôle de validité et de formaliser la récupération.

L'information de type « expertise »

Ce type d'information concerne les experts de l'entreprise ou en relation avec elle, leur localisation, leur connaissance, leur talent et rassemble les savoir-faire. Il s'agit de la mémoire de l'entreprise.

L'information de type « foires et salons »

Ce type d'information revêt une grande importance pour l'entreprise car dans un même lieu sont rassemblés les concurrents et les clients. Il est nécessaire d'organiser une stratégie de recueil des informations pertinentes : visites de stands, questions à poser, plaquettes à récupérer et de formaliser les informations ainsi recueillies.

L'information de type « floue » ou informelle revêt une importance stratégique indéniable et représente la cible privilégiée des managers en raison de sa spontanéité. L'information formalisée serait pour eux trop générale et trop tardive (manque de « fraîcheur »). Elle offrirait une description incomplète de l'environnement externe de l'entreprise.

1.3.2 Typologie des sources d'information

On peut catégoriser ces sources d'informations selon plusieurs critères différents : le coût d'accès à l'information (sources payantes ou gratuites), la fiabilité de l'information, sources en libre accès ou sur abonnement, la structure du contenu (sources qui donnent accès à une information structurée ou non), la mise à jour des informations, et le volume d'informations disponibles, l'origine des sources, avec un résumé ou non, sources d'informations sous forme électronique ou

papier, …et aussi la nature du document, la langue.

	Fiabilité	**Structure**	**Origine**	**Coût**	**Mise à jour**	**Volumes**
Internet	non	non	externe	faible	constante	immense
CD-ROM	oui	oui	externe	faible	non	limité
Intranet	?	non	interne	nul	?	limité
Bases de données en ligne	oui	oui	externe	élevé	irrégulière	élevé
Bases de données locales	?	oui	interne	nul	?	limité

Table 1 : Caractérisation des sources d'information électroniques.

La détermination du type et de la nature de ressources est très importante dans le processus de veille. Dans le cadre de la thèse, nous nous intéressons plutôt aux sources électroniques. Voici une table qui propose une caractérisation des cinq sources d'information électroniques principales, selon ces critères :

Les points d'interrogation signifient qu'il est difficile de caractériser les informations qui sont internes à l'entreprise, car selon les cas elles sont plus ou moins contrôlées (donc fiables), et plus ou moins mises à jour.

2 Web Sémantique et application à la recherche d'information

Proclamé prochaine évolution du web, le Web sémantique a attiré depuis 1999 l'attention de nombreux chercheurs. Il s'agit d'arriver à un web « intelligent », où les informations ne seraient plus stockées mais « comprises » par les ordinateurs afin d'apporter à l'utilisateur ce qu'il cherche vraiment. D'un certain point de vue, le web sémantique est une évolution pour les systèmes de recherche d'information. Afin d'offrir la capacité de traitement automatique sur des documents non structurés, le web sémantique, en ajoutant aux informations existantes une couche de métadonnées, les rend exploitables par les ordinateurs. Ces métadonnées apporteront des sémantiques sans ambiguïté pour automatiser les traitements. Ainsi, une fois mis en place, le web sémantique enrichira l'exploration d'informations sur les moteurs de recherche. Ce chapitre présente un état de l'art sur les travaux autour du web sémantique. Après un bref aperçu sur la vision du web sémantique, nous étudierons ses composants principaux, ses langages de différents niveaux pour la représentation de connaissances et ses domaines d'application.

2.1 Web Sémantique

La notion de web sémantique fait référence à la vision du web de demain dans lequel les utilisateurs devraient être déchargés d'une bonne partie de leurs tâches de recherche et d'exploitation des résultats, grâce aux capacités accrues des machines à accéder aux contenus des ressources et à effectuer des raisonnements sur ceux-ci.

Concrètement, le web sémantique est une infrastructure qui permet l'utilisation de connaissances formalisées en plus du contenu informel que l'on peut trouver dans le web actuel. Cette infrastructure s'appuie sur un certain niveau de consensus portant, par exemple, sur les langages de représentation ou sur les ontologies utilisées. Ainsi, elle permet, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ontologies. Grâce à la formalisation de connaissances, elle peut faciliter la mise en oeuvre de calculs et de raisonnements complexes tout en offrant des garanties supérieures sur leur validité. Mais restreindre le web sémantique à cette infrastructure serait trop limitatif. Sur la base de sémantiques bien définies pour ses ressources, le web sémantique pourra fournir aux utilisateurs, par le moyen d'agents logiciels, des services automatiques et avancés [Laublet et al., 2002]. Comme l'écrivent en substance [Berners-Lee et al., 2001], « le web sémantique est une extension du web courant, dans laquelle on donne à une information un sens bien défini pour permettre aux ordinateurs et aux personnes de travailler en coopération ».

L'architecture du web sémantique s'appuie sur une pyramide de langages proposée par Tim Berners-Lee pour représenter des connaissances sur le web en satisfaisant les critères de standardisation, d'interopérabilité et de flexibilité. Cette architecture en couches (Figure 3) peut permettre une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs. Un langage de la couche haute doit être une extension du langage de la couche au-dessous. Aujourd'hui seules les couches basses sont relativement stabilisées. La liste suivante introduit la fonction principale de chaque couche dans l'architecture du web sémantique :

- XML est utilisé comme couche de base syntaxique du web sémantique. Le langage XML est actuellement considéré comme un standard pour le transport de données sur le web.
- La couche RDF représente les métadonnées pour les ressources web.
- La couche « ontologie », fondée sur une formalisation commune, spécifie la sémantique de métadonnées fournies dans le web sémantique.
- La couche « logique » s'appuie sur des règles d'inférence qui permettent le raisonnement intelligent exécuté par des agents logiciels.

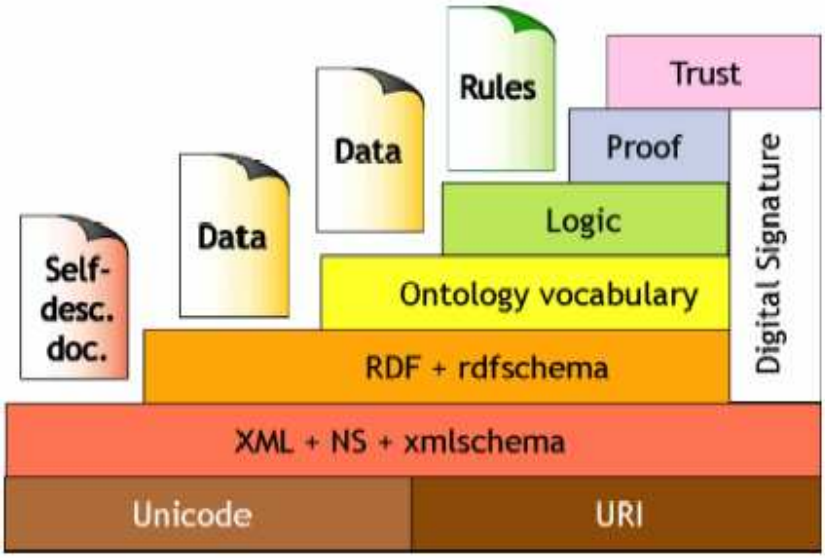


Figure 3 Les couches du Web Sémantique

2.2 Les principales composantes du web sémantique

2.2.1 Ontologie

Une ontologie, en philosophie, est une théorie à propos de la nature de l'existant, des types de choses qui existent et des types de leurs liens. Les chercheurs du monde de l'intelligence artificielle ont adopté ce terme dans leur propre jargon, et pour eux une ontologie est la spécification explicite d'une conceptualisation partagée qui présente une vue du monde réel dans un domaine spécifique.

Le but de l'ontologie, selon [Uschold et Jasper, 1999] est :

- La communication (entre humains et/ou organisations) : Le bénéfice de l'usage d'ontologie ici, est sans ambiguïté. Dans l'ontologie, il n'y a jamais deux termes ayant la même sémantique. Cette situation se produit souvent, au contraire, si l'on utilise un langage naturel pour la communication.
- L'interopérabilité (machines et systèmes) : L'ontologie peut être utilisée comme un modèle intermédiaire pour la traduction entre les modélisations de différentes collections d'objets. L'ontologie sert à définir le format d'échange entre les systèmes.
- L'ingénierie des systèmes : L'ontologie peut servir divers aspects du développement des systèmes d'informations. Premièrement, elle peut assister le processus de construction de spécification de système. L'usage d'ontologie rend les documents du processus plus compréhensibles, évite l'ambiguïté dans la spécification. En outre, une représentation formelle d'ontologie permet un traitement automatique du développement. Elle soutient également l'automatisation du processus de vérification de la fiabilité des systèmes.

2.2.1.1 Quelle définition?

La définition la plus célèbre est celle de [Gruber, 1993] « une ontologie est une spécification explicite d'une conceptualisation » (An ontology is an explicit specification of a conceptualization).

N. Guarino affine la définition de T. Gruber en considérant les ontologies comme des spécifications partielles et formelles d'une conceptualisation [Guarino et Giaretta, 1995]. Les ontologies sont formelles car exprimées sous forme logique, et partielles car une conceptualisation ne peut pas toujours être entièrement formalisée dans un cadre logique, du fait d'ambiguïtés ou du fait qu'aucune représentation de leur sémantique n'existe dans le langage de représentation d'ontologies choisi.

Pour [Swartout, 1996] une ontologie est un ensemble structuré de termes décrivant un domaine et utilisable comme un noyau d'une base de connaissances.

[Bachimont, 2000] présente une ontologie comme le résultat d'une modélisation. Celle-ci porte sur la caractérisation de primitives pour la représentation des connaissances.

Les auteurs de [Bouquet et al., 2003] définissent les ontologies comme « des modèles partagés d'un domaine encodant une vue qui est commune à un ensemble de différentes parties ».

De nombreuses définitions offrent des points de vues divers et complémentaires.

2.2.1.2 Classification des ontologies

La classification de [van Heijst et al., 1997] repose sur deux critères :

- la structure de la conceptualisation,
- le sujet de la conceptualisation.

Pour le premier critère, van Heijst et ses collègues distinguent trois catégories, à savoir (i) les ontologies terminologiques (lexiques, glossaires…), (ii) les ontologies d'information (schéma d'une BD) et (iii) les ontologies des modèles de connaissances.

En ce qui concerne le sujet de la conceptualisation, ils distinguent quatre catégories :

- *Les ontologies d'application* : contiennent toutes les informations nécessaires pour modéliser les connaissances pour une application particulière.
- *Les ontologies de domaine* : fournissent un ensemble de concepts et de relations décrivant les connaissances d'un domaine spécifique.
- *Les ontologies génériques* (dites aussi de haut niveau) : similaires aux ontologies de domaine, mais les concepts qui y sont définis sont plus génériques et décrivent des connaissances tels que l'état, l'action, l'espace et les composants. Généralement les concepts d'une ontologie domaine sont des spécialisations des concepts d'une ontologie de haut niveau.
- *Les ontologies de représentation ou méta-ontologies* : fournissent des primitives de formalisation pour la représentation des connaissances. Elles

sont généralement utilisées pour écrire les ontologies de domaine et les ontologies de haut niveau.

2.2.1.3 Le cycle de vie des ontologies

Les ontologies étant destinées à être utilisées comme des composants logiciels dans des systèmes répondant à des objectifs opérationnels différents, leur développement doit s'appuyer sur les mêmes principes que ceux appliqués en génie logiciel. En particulier, les ontologies doivent être considérées comme des objets techniques évolutifs et possédant un cycle de vie qui nécessite d'être spécifié. Les activités liées aux ontologies sont d'une part des activités de gestion de projet (planification, contrôle, assurance qualité), et d'autre part des activités de développement (spécification, conceptualisation, formalisation) ; s'y ajoutent des activités transversales de support telles que l'évaluation, la documentation, la gestion de la configuration [Blazquez et al., 1998]. Un cycle de vie inspiré du génie logiciel est proposé dans [Dieng et al., 2005]. Il comprend une étape initiale d'évaluation des besoins, une étape de construction, une étape de diffusion, et une étape d'utilisation. Après chaque utilisation significative, l'ontologie et les besoins sont réévalués et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite.

La phase de construction peut être décomposée en 3 étapes : conceptualisation, ontologisation, opérationalisation. L'étape d'ontologisation peut être complétée d'une étape d'intégration au cours de laquelle une ou plusieurs ontologies vont être importées dans l'ontologie à construire [Fernandez et al., 1997]. Fernandez insiste sur le fait que les activités de documentation et d'évaluation sont nécessaires à chaque étape du processus de construction, l'évaluation précoce permettant de limiter la propagation d'erreurs. Le processus de construction peut être intégré au cycle de vie d'une ontologie comme indiqué en figure 4 [Gandon, 2002]. La section suivante va être plus spécifiquement consacrée aux méthodologies mises en oeuvre lors de la phase de construction afin, en particulier, de guider les choix délicats de conceptualisation et de représentation.

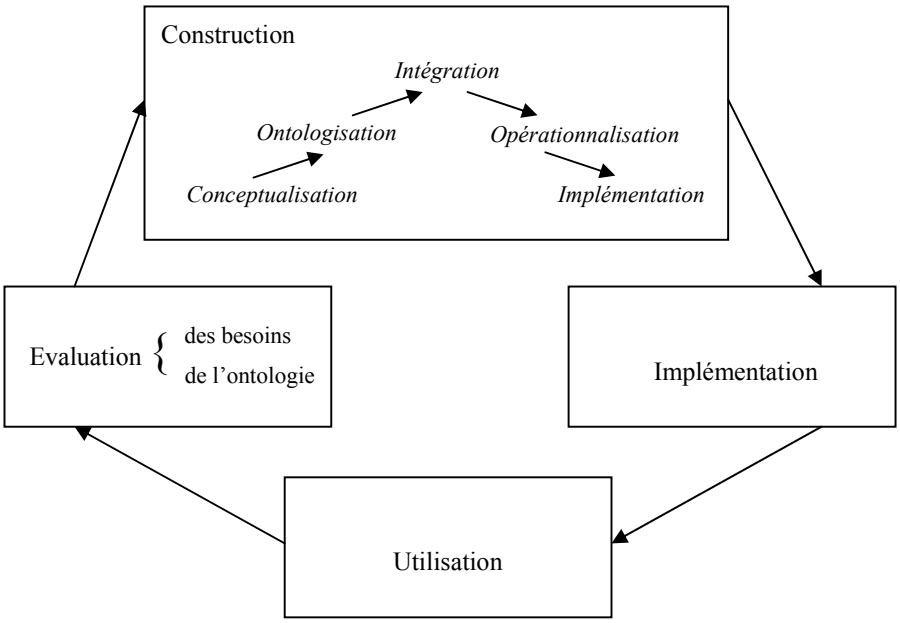


Figure 4 Le cycle de vie d'une ontologie

2.2.1.4 La construction d'une ontologie

Bien qu'aucune méthodologie générale n'ait pour l'instant réussi à s'imposer, de nombreux principes et critères de construction d'ontologies ont été proposés. Ces méthodologies peuvent porter sur l'ensemble du processus et guider l'ontologiste à toutes les étapes de la construction.

Les trois méthodologies les plus souvents citées sont :

- La Methontology de [Gomez-Perez, 1998] et [Fernandez et al., 1997], qui couvre tout le cycle de vie d'une ontologie
- La méthodologie de Uschold et King [Uschold et al., 1995], élaborée à la suite de l'expérience de la construction de l'Enterprise Ontology
- La méthodologie de Gruninger et Fox [Gruninger et Fox, 1995], basée sur l'expérience de la construction d'une ontologie dans le cadre du projet Tove.

D'autres méthodologies se focalisent sur une des étapes du processus de représentation des connaissances. Celle présentée dans [Aussenac-Gilles et al., 2000] insiste sur l'étape de conceptualisation par l'analyse d'un corpus textuel. La méthodologie Ontospec de G. Kassel constitue une aide à la structuration des

hiérarchies de concepts et de relations durant la phase d'ontologisation [Kassel, 2002]. C'est également le cas des principes différentiels énoncés par B. Bachimont [Bachimont, 2001] et des critères de classification des propriétés, proposés par N. Guarino et C. Welty [Guarino et Welty, 2000]. La méthode On-to-knowledge [Sure et al.,1999], a été mise en œuvre pour répondre à la problématique du développement d'ontologies dans le cadre du Web sémantique. Mais quelle que soit la méthodologie adoptée, le processus de construction d'une ontologie est une collaboration qui réunit des experts du domaine de connaissance, des ingénieurs de la connaissance, même parfois les futurs utilisateurs de l'ontologie [Farquhar et al., 2000]. Cette collaboration ne peut être fructueuse que si les objectifs du processus ont été clairement définis, ainsi que les besoins qui en découlent.

Les environnements de développement d'ontologies contiennent généralement un ensemble d'outils, centrés autour d'un éditeur de connaissances et supposés aider à la construction d'une ontologie ainsi qu'à sa formalisation dans un langage tel que RDFS ou OWL. Protégé [Noy et al., 2001] est l'éditeur de connaissances le plus connu à l'heure actuelle, avec une architecture extensible, permettant d'intégrer des greffons (pluggins) afin d'étendre l'éditeur pour de nouvelles fonctionnalités et la prise en charge de nouveaux langages. Ontoedit [Sure et al., 2002] est un éditeur de connaissances fondé sur un processus de développement d'ontologies suivant les différentes étapes de la méthode On-To-Knowledge. Webode [Arpirez et al., 2003] est quant à lui un environnement de développement fondé sur une architecture client-serveur et offrant des outils couvrant l'ensemble du cycle de vie d'une ontologie, selon la méthode Methontology.

2.2.2 Annotation sémantique

Pour le Web Sémantique, l'un des aspects les plus importants est de pouvoir manipuler des annotations sémantiques de documents Web, puisque le Web Sémantique permettra aux machines de comprendre la sémantique des documents et des données. Les annotations sémantiques décrivent le contenu des documents, en associant une sémantique à ces descriptions. On peut les considérer comme des

méta-données de documents, ressources du Web.

Clairement, la sémantique de l'annotation est fondée sur des vocabulaires dans les ontologies qui sont spécifiées explicitement dans un langage de représentation.

2.2.3 Langage de représentation de connaissance

2.2.3.1 XML

XML est un méta-langage proposé par le W3C permettant de représentant un document texte de manière arborescente en utilisant un système de balisage.

Ce langage a été élaboré pour faciliter l'échange, le partage et la publication des données à travers le web. Ainsi, la majorité des langages/modèles proposés pour le web sémantique sont exprimés en XML.

XML permet de structurer un document en définissant ses propres balises en fonction des besoins et sans tenir compte ni de la signification de cette structure ni des systèmes informatiques qui vont l'exploiter. Des standards comme XPath [clark 99] et XQuery [Boag 2004] ont été développés afin de parcourir et d'interroger l'arborescence XML des documents.

2.2.3.2 RDF/RDFS

RDF (Resource Description FrameWork) est la spécification d'un système d'expression d'assertions sémantiques simples. Ce système est une recommandation du World Wide Web consortium (W3C). La syntaxe de RDF est basée sur celle de XML. Le modèle de base de RDF est conçu pour permettre d'associer des attributs aux ressources du Web en utilisant la description de méta-données sémantiques. Les objectifs initiaux de RDF étaient la représentation et une meilleure exploitation des méta-données. Mais, de manière plus générale, RDF permet de voir le Web comme un ensemble de ressources reliées par les liens étiquetés « sémantiquement».

Une assertion RDF est donnée par un triplet (Ressource, Propriété, Valeur).

- Ressource : toute expression RDF a pour but de décrire une ressource. Il s'agit d'une entité référencée par un identificateur unique. Les identificateurs uniques du Web sont les URIs. Les ressources sont variées :

une page Web, une partie d'une page Web, un site Web complet, un objet non accessible par le Web comme un livre…

- Propriété : un aspect spécifique, un attribut, une caractéristique ou relation utilisée pour décrire une ressource...
- Valeur : Un littéral (simple chaîne de caractères) ou une ressource.

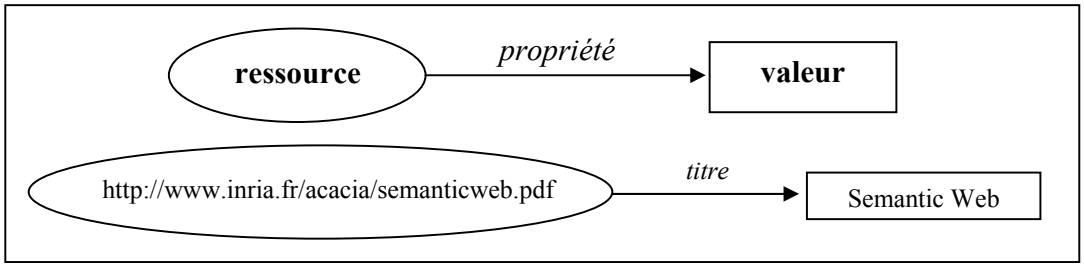


Figure 5 Exemple d'un modèle RDF

RDFS [Lassila and Swick, 1999] est un méta modèle proposé par le W3C en tant que recommandation afin de permettre la définition de schéma/modèle décrivant l'univers sémantique des déclarations RDF.

RDFS fournit ainsi un système de typage pour les déclarations RDF. Il permet la définition des classes et des sous-classes (rdfs:Class, rdfs:subClassOf) décrivant les ressources à annoter et donnant un sens aux propriétés associées aux ressources. Il permet aussi la formulation de contraintes sur les valeurs associées à une propriété afin de lui assurer une signification (rdfs:domain, rdfs:range).

Dans le contexte du web sémantique, RDFS est utilisé pour formaliser les ontologies sur lesquelles vont se baser les annotations RDF. Dans l'exemple suivant, nous employons des primitives du schéma RDF pour définir la classe « rapport de recherche » et décrire une des caractéristiques des ressources appartenant à cette classe, telle que leur titre.

```
<rdf:RDF
  xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:rdfs=http://www.w3.org/2000/01/rdf-schema#>

  <rdfs:Class rdf:ID="ResearchReport">
    <rdfs:subClassOf rdf:resource="#Report"/>
    <rdfs:comment xml:lang="en">Report presenting studies.</rdfs:comment>
      <rdfs:comment        xml:lang="fr">Rapport        presentant        des
      etudes.</rdfs:comment>
      <rdfs:label xml:lang="en">research report</rdfs:label>
```

```
  <rdfs:label xml:lang="fr">rapport de recherche</rdfs:label>
</rdfs:Class>
<rdf:Property rdf:ID="Title">
 <rdfs:range rdf:resource="&rdfs;Literal"/>
 <rdfs:domain rdf:resource="#Report"/>
 <rdfs:label xml:lang="en">title</rdfs:label>
 <rdfs:label xml:lang="fr">titre</rdfs:label>
</rdf:Property>
</rdf :RDF>
```

Avec les primitives rdfs :subClassOf et rdfs :subRelationOf, nous pourrons définir des liens de « spécialisation » qui permettent aux classes et aux relations d'hériter des caractéristiques définies dans des classes ou des relations parentes.

2.2.3.3 DAML+OIL

DAML+OIL est un langage de balises pour la représentation et l'échange d'ontologies. C'est le résultat de la fusion des langages OIL [Fensel et al, 2001] et DAML-ONT (DARPA Agent Markup Language)[DAML, 2000] et il s'appuie sur les langages RDF et RDF Schema en les enrichissant avec de nouvelles primitives. D'un point de vue formel, DAML + OIL peut être vu comme une logique de description très expressive. L'expressivité du langage est déterminée par les types de constructeurs supportés, c'est à dire, ceux qui permettent de définir des classes et des propriétés ainsi que des axiomes.

2.2.3.4 OWL

Le W3C cherche à proposer un standard, connu actuellement sous le nom de OWL (Ontology Web Language [Dean et al., 2003]), dérivé de DAML+OIL [Horrocks et al., 2001], un langage qui s'appuie sur la logique de description. Le langage OWL devrait être construit sur RDFS tout en disposant d'une syntaxe XML. Sans être exhaustif, disons au moins qu'il permet la possibilité de définir des classes de manière plus complexe : celles correspondant aux connecteurs de la logique de description (intersection, union, restriction, etc.), des propriétés inverses ou transitives ou bien encore des restrictions de cardinalité sur les propriétés. En se fondant sur une logique de description, un tel langage a une sémantique formelle

claire, ce qui permet de le doter de services inférentiels [Laublet et al., 2002].

La compatibilité avec RDF(S) et l'importante expressivité de OWL posent par ailleurs d'autres problèmes. Tout d'abord, certaines constructions de RDF(S) impliquent que les mécanismes de raisonnements associés au langage soient indécidables. L'aspect réflexif de RDF(S), qui permet de redéfinir les éléments du langage, empêche, de plus, la mise en œuvre d'une sémantique standard et fixée de manière définitive [Horrocks et al., 2003]. En outre, même si ces constructions particulières à RDF(S) ne sont pas utilisées, la Logique de Description (LD) sous-jacente à OWL reste très expressive et les inférences associées sont de fait très complexes. Pour ces raisons, le langage OWL est décomposé en trois sous-langages : OWL Lite, OWL DL et OWL Full, qui sont conçus comme trois couches successives, du plus simple (OWL Lite) au plus complexe (OWL Full).

- OWL Full correspond au langage OWL dans son ensemble, c'est-à-dire qui inclut tous les constructeurs de OWL et toutes les possibilités de RDF(S). L'inférence en OWL Full constitue ainsi un problème indécidable.
- OWL DL est un sous-langage de OWL Full correspondant à la LD SHOIN (D). OWL ne permet pas l'utilisation de certaines constructions particulières de RDF(S) (réflexivité, etc).
- OWL Lite est un sous-langage de OWL correspondant lui aussi à une LD, mais moins expressive. OWL Lite ne reprend de OWL DL que les constructeurs jugés les plus utiles et les plus faciles à mettre en œuvre.

2.2.4 Système et outils d'annotation

2.2.4.1 Annotation manuelle

La plupart des outils d'annotation de base permettent à leurs utilisateurs de créer manuellement des annotations. Ils ont beaucoup en commun avec les outils purement textuels d'annotation mais fournissent un certain support pour des ontologies. Par exemple, le navigateur et l'éditeur Web Amaya du W3C (Quint et Vatton 1997) peuvent baliser des documents Web en XML ou HTML.

Annotea est un système promu par le W3C permettant de faire des annotations de documents Web dans un environnement collaboratif. Ces annotations peuvent

être des commentaires, des explications correspondant à un document entier ou à des fragments d'un document. Chaque fois qu'un utilisateur accède à un document, il a aussi accès à toutes les annotations liées à ce document. Les annotations sont stockées sur un serveur (ZAnnot server) comme des métadonnées et présentées aux utilisateurs par des clients.

COHSE Annotator (Bechhofer et Goble, 2001) produit les annotations qui sont compatibles avec des normes d'Annotea, bien que les annotations soient conçues comme des hyperliens enregistrés en utilisant le service de liens distribué (Carr et al., 1995). Dans ce scénario, les hyperliens automatiquement appliqués sont acceptables mais seule un service de « word-matching» qui met en surbrillance des termes d'ontologie dans le texte a été mis en application pour l'instant.

Le système Mangrove est un autre exemple de système d'annotation manuelle convivial pour l'utilisateur (McDowell et al.. 2003). Le but du système était de permettre aux utilisateurs de baliser leurs documents HTML en utilisant les données créées par un certain nombre de services sémantiques tels qu'un calendrier des événements. L'outil d'annotation offre une interface utilisateur qui permet à des utilisateurs d'associer une sélection d'étiquettes au texte qu'ils mettent en surbrillance.

Quelques outils manuels d'annotation ont été développés pour fournir une assistance aux utilisateurs plus avancés et un degré d'automatisation (partielle ou totale) de l'annotation. OntoMat Annotizer est un outil permettant de produire des annotations selon les principes de la plate-forme CREAM. Il offre un navigateur Web pour afficher la page annotée et fournit quelques fonctions utilisateur conviviales pour l'annotation manuelle.

OntoMat a été étendu pour inclure l'aide à l'annotation semi-automatique. La première de ces extensions est l'outil S-CREAM, (Handschuh et al.2003), qui utilise le système d'extraction d'information (IE) Amilcare (Ciravegna et Wilks 2003). L'utilisateur annote un document servant d'exemple et le système apprend comment reproduire l'annotation de l'utilisateur, pour pouvoir suggérer des annotations pour de nouveaux documents. OntoMat incorpore également des méthodes pour l'annotation profonde (Volz et al.2004), c.-à-d. l'annotation pour les Pages Web qui

sont produites à partir des bases de données. Une version commerciale d'OntoMat, appelée OntoAnnotate5, est commercialisée par Ontoprise

2.2.5 Méthodes et outils d'extraction d'information pour l'annotation automatique

La majeure partie de la technologie actuelle est basée sur de l'annotation humaine (manuelle). Mais l'annotation manuelle est répétitive, coûteuse, et consomme du temps. Convaincre les auteurs d'annoter les documents du Web en utilisant des ontologies est difficile et nécessiterait une action au niveau mondial. D'ailleurs, l'annotation statique associée à un document peut : (1) être inachevée ou incorrecte quand le créateur n'est pas assez habile ; (2) devenir désuète, c.-à-d. ne pas prendre en compte les mises à jour des pages. Donc, il est important de disposer de méthodes pour l'annotation automatique des pages. L'annotation initiale liée au document perd son importance parce qu'à tout moment, il est possible de ré-annoter automatiquement le document. Dans cette section nous allons résumer les méthodes et les systèmes d'annotation automatique les plus courants.

En fonction des types de textes que distinguent les approches d'extraction d'information pour le traitement des informations textuelles, les techniques de traitement automatique de la langue naturelle (TALN) jouent un rôle primordial. Mais ces techniques ne sont plus suffisantes pour extraire des données structurées, car les informations ne gardent plus toutes leurs caractéristiques linguistiques (comme leurs informations grammaticales). Par exemple dans une page Web bien structurée, certaines parties de texte ne comprennent pas des phrases complètes.

Dans cette section nous considérons les deux types d'outils d'annotation ceux qui incluent les composants d'automatisation et fournissent des suggestions pour des annotations, mais exigent toujours l'intervention par des utilisateurs, et ceux qui saisissent des annotations automatiquement sur une grande échelle. Certains sont encore limités à l'utilisation par des spécialistes tandis que d'autres conviennent aux utilisateurs non spécialistes (travailleurs intellectuels). Les systèmes automatisés, destinés à aider des travailleurs du savoir (knowledge workers) doivent dans leur interface utilisateur minimiser l'intrusion tout en maximisant l'exactitude.

L'automatisation peut généralement être considérée comme tombant dans trois catégories. Le système le plus fondamental utilise des règles ou des «wrappers» qui essayent de capturer les motifs ou patrons connus pour les annotations. Alors il existe deux genres de systèmes qui apprennent comment annoter. Les systèmes supervisés apprennent des exemples d'annotations balisés par l'utilisateur. Un problème avec ces méthodes est que la sélection d'assez bons exemples est une tâche non triviale. Afin de résoudre ce problème les systèmes non supervisés utilisent une variété de stratégies pour apprendre comment annoter sans supervision d'utilisateur, mais leur exactitude est encore limitée.

2.2.5.1 L'extraction adaptative d'information

Les systèmes adaptatifs d'extraction d'information utilisent l'apprentissage automatique (« machine learning ») pour apprendre comment s'adapter à une nouvelle application ou un nouveau domaine [Kushmerick, 2002]. L'idée générale est la suivante : un expert humain annote un corpus de documents en spécifiant des fragments d'information à extraire dans chaque document. Un algorithme d'apprentissage sera appliqué pour entraîner le système, sur un ensemble d'exemples, puis le système d'apprentissage généralise à partir de ces exemples afin de produire des règles permettant d'extraire correctement des fragments d'information de même nature (chiffre, nom propre, date, patrons particuliers du domaine, etc..) dans d'autres nouveaux documents.

2.2.5.2 Annotation automatique pour texte libre

Jusqu'à présent, il existe peu de systèmes d'annotation automatique de texte libre. Les systèmes les plus connus sont : MnM [Vargas,2002] (Open University), Ontomat [10] (University of Karlsruhe), KIM Platform [Popov,2003] (Ontotext Lab), PIA-Core [Collier,2002](PIA project). Ils partagent des caractéristiques communes :

- Intégration avec des ontologies (pour insérer dans le texte des concepts, utilisés comme des balises, pour annoter une partie de texte)

- Le module d'extraction d'information adaptative repose sur des outils linguistiques : MnM et Ontomat utilisent directement Amilcare et KIM utilise GATE [Gate].

MnM a été conçu pour baliser des données d'apprentissage pour des outils d'extraction d'information (EI) plutôt qu'un outil d'annotation intrinsèquement (Vargas-Vera et al.2003). Ceci signifie qu'il stocke des documents balisés en tant que versions étiquetées de l'original, plutôt que les formats RDF employés par la communauté Web Sémantique. Il fournit une assistance raisonnable aux utilisateurs avec un navigateur HTML pour afficher les documents et avec le fonctionnement d'un navigateur d'ontologie. Une force de MnM est qu'il fournit des APIs ouvertes pour se connecter aux serveurs d'ontologie et pour intégrer les outils d'extraction de l'information, ce qui le rend flexible au sujet des formats et des méthodes qu'il utilise.

Melita (Ciravegna et al.2002) est un outil sémantique automatisé d'annotation guidé par l'utilisateur qui rend deux stratégies principales disponibles à l'utilisateur. D'une part, il fournit un système adaptatif d'extraction de l'information (Amilcare) qui apprend comment annoter les documents par la généralisation sur les annotations d'utilisateur. L'annotation est donc un processus qui commence par exiger l'annotation complète de l'utilisateur aux premières étapes, mais termine en faisant simplement vérifier par l'utilisateur l'exactitude des suggestions faites par le système. D'autre part, il fournit des fonctionnements pour l'écriture de règles (basées sur des expressions régulières) pour permettre aux utilisateurs avancés de définir leurs propres règles. Dans Melita, les documents ne sont pas choisis aléatoirement pour l'annotation, mais sont plutôt choisis automatiquement selon l'utilité souhaitée, pour le système d'EI. Le système d'EI d'Amilcare a été incorporé dans K@, un système légal de gestion des connaissances avec des capacités sémantiques basées sur RDF produites par Quinary (Gilardoni et al 2005).

CAFETIERE est un système basé sur les règles pour générer des annotations XML. Il a été développé dans le cadre du projet Parmenides (Black et al. 2005). Il a été utilisé, par exemple, pour annoter le corpus biomédical GENIA (Rinaldi et al. 2004). Les techniques de fouille de textes sont utilisées pour suggérer des

annotations aux analystes (Vasilakopoulos 2004). Le projet Parmenides a également expérimenté une approche « clustering » pour suggérer des concepts et des relations pour étendre des ontologies (Sipiliopoulou et al. 2004).

Armadillo est un système pour la création non supervisée des bases de connaissances à partir de grands entrepôts (par exemple le Web) aussi bien que l'annotation de documents (Ciravegna et al. 2004). Il utilise la redondance d'information dans des entrepôts pour amorcer l'apprentissage d'une poignée d'exemples choisis par l'utilisateur. Alors l'extraction d'information adaptative est utilisée pour généraliser ces exemples et pour trouver de nouveaux faits. La confirmation par plusieurs sources (par exemple documents) est alors exigée pour vérifier la qualité des données saisies. Après confirmation, l'apprentissage peut être lancé une nouvelle fois. Ce processus peut être répété jusqu'à ce que l'utilisateur soit satisfait de la qualité d'information issue après l'apprentissage. Armadillo utilise un certain nombre de techniques, à partir des recherches basées sur des mot-clés jusqu'à l'extraction de l'information adaptative pour l'intégration de l'information.

L'algorithme d'apprentissage appliqué dans Amicare [Ciravegna,2002] (et donc dans Ontomat et MnM) est LP2, une technique d'apprentissage relationnel ; plus précisément c'est un algorithme « Sequential Covering » qui utilise des informations linguistiques dans le processus d'apprentissage. Après l'évaluation, le taux de succès est de 70% sur les documents texte.

KnowItAll (Etzioni et al.2005) automatise l'extraction de grandes bases de faits à partir du Web dans un mode similaire à Armadillo. La différence la plus notable est la façon dont le système évalue la plausibilité des candidats extraits. Ceci est fait en utilisant la mesure de PMI (pointwise mutual information) plutôt que de donner des poids provenant de preuves multiples à partir des oracles spécifiques du domaine. La mesure de PMI est le rapport entre le nombre de résultats obtenus par le moteur de recherche avec une requête sous forme d'une phrase discriminative (par exemple "Liège est une ville") et le nombre de résultats obtenus avec une requête sous forme de fait extrait (par exemple "Liège"). En outre, KnowItAll n'exige pas de connaissances initiales. Les auteurs ont fourni trois extensions au

système (apprentissage de pattern, extraction de sous-classe et extraction de liste) pour améliorer la performance globale.

Une autre approche à l'apprentissage des annotations qui exploite une petite partie du Web est PANKOW (Pattern-based Annotation through Knowledge On the Web) (Cimiano et al. 2004). PANKOW utilise un intervalle de patrons relativement rares, mais informatifs, syntaxiques pour baliser des phrases candidates situées dans des pages Web, sans devoir manuellement produire un ensemble initial de pages Web balisées et alors passer à une étape d'apprentissage supervisé.

SemTag est un autre exemple d'un outil qui se concentre uniquement sur l'annotation automatique (Dill et al. 2003). Il est basé sur la plate-forme d'analyse des textes Seeker d'IBM et emploie des fonctions de similarité pour identifier les entités qui apparaissent dans les contextes similaires afin de baliser des exemples. Le problème principal du balisage automatique à grande échelle est l'ambiguïté, par exemple une chaîne de caractères identiques, telle que "Niger", peut se rapporter à différentes choses : un fleuve ou un pays. On propose un algorithme de désambiguïsation basé sur une taxonomie (TBD) pour résoudre ce problème. Cet outil est destiné à des spécialistes plutôt qu'à des travailleurs du savoir (knowledge worker).

KIM (Popov et al. 2003) (Popov et al. 2004) utilise des techniques d'extraction de l'information pour construire une grande base de connaissances des annotations. Les annotations dans KIM sont des méta-données sous forme d'entités nommées (les gens, les endroits etc...) qui sont définies dans l'ontologie KIMO et sont identifiées principalement à partir de la référence à des « gazetteers » (les outils effectuant la correspondance entre les termes dans un texte et le vocabulaire d'un dictionnaires) extrêmement grands. C'est restrictif, et ce serait un défi significatif de recherches pour étendre la méthodologie de KIM à des ontologies spécifiques au domaine. Toutefois les entités nommées sont une classe de méta-donnée avec une large utilisation. Par l'exemple, dans l'application Rich News, KIM a été utilisé pour aider à annoter des actualités de télévision et de radio en exploitant le fait que des actualités Web sur les mêmes sujets sont souvent publiées en parallèle (Dowman et Al 2005). La plate-forme de KIM est bien placée pour présenter les

genres de services d'analyse de données et de recherche qui peuvent être fournis avec de grandes bases de connaissance des annotations.

Une importante limite de tous ces systèmes est que les annotations reposent seulement sur les « Named Entity » (NE) apparaissant dans le texte, c'est-à-dire qu'on n'annote que des types sémantiques spécifiques : personne, organisation, adresse, date, coût, nombre, etc. Donc les annotations sémantiques sont ici des méta-données qui associent des entités du texte à leur description. Les techniques linguistiques sont exploitées pour reconnaître les NE.

L'application de l'apprentissage automatique pour de l'annotation adaptative dans PIA-Core est en cours de développement (les approches SVM et HMM sont testées mais il n'y pas de résultats annoncés).

Dans ces systèmes, l'ontologie n'aide en rien au processus d'apprentissage. Elle joue seulement un rôle de fournisseur des connaissances du domaine (pour mettre des « balises » autour des informations extraites). Les termes dans le texte qui correspondent aux concepts de l'ontologie ne sont pas reconnus (sauf les NE).

2.2.5.3 Annotation automatique pour données semi-structurées

Certains systèmes connus, appelés « Wrapper Induction », appliquent des techniques d'apprentissage automatique. Il y a deux approches générales :

Approches « aux états finis » (finite state) : dans ces approches, l'algorithme d'apprentissage apprend « les connaissances d'extraction » sous forme d'automates. Par exemple : algorithme LR et ses dérivés, modèles HMM.

Approches « relationnelles » : dans ces approches, l'algorithme d'apprentissage apprend « les connaissances d'extraction » sous forme de programmes logiques « Prolog-like ».

Une limite de tous ces algorithmes est qu'ils traitent des données semi-structurées comme du texte uni et qu'ils cherchent à trouver des délimiteurs pour extraire des informations. La différence avec les algorithmes pour le texte libre est qu'ils évitent d'appliquer des analyses linguistiques.

Lixto est un système d'extraction de l'information du Web permettant à des « wrappers » d'être définis pour convertir les ressources non structurées en des

ressources structurées. L'outil permet à des utilisateurs de créer des wrappers en mode interactif et visuellement en choisissant des morceaux pertinents d'information (Baumgartner et al.2001). Il a été initialement développé à l'université technique de Vienne.

2.3 Système multi-agents de recherche d'information

Depuis une dizaine d'années, les systèmes multi-agents ont connu un grand essor et sont appliqués à des domaines très variés comme, par exemple, le domaine de la simulation et de la vie artificielle, la robotique, le traitement d'images, la recherche d'information.

Les systèmes multi-agents sont issus de l'intelligence artificielle distribuée (IAD), une branche de l'intelligence artificielle qui s'articule autour de trois axes :

- La résolution distribuée des problèmes qui s'intéresse à la manière de diviser un problème en un ensemble d'entités distribuées et coopérantes et à la manière de partager la connaissance du problème afin d'en obtenir la solution.

- L'intelligence artificielle parallèle qui développe des langages et des algorithmes parallèles pour l'intelligence artificielle (IA) visant ainsi l'amélioration des performances des systèmes d'IA.

- Les systèmes multi-agents qui privilégient une approche décentralisée de la modélisation et mettent l'accent sur les aspects collectifs des systèmes.

Les systèmes multi-agents, comme leur nom l'indique, sont des systèmes constitués de plusieurs agents. Il est donc nécessaire de commencer par définir ce que l'on appelle un agent. La notion d'agent définie, nous introduirons le concept de système multi-agents.

2.3.1 Notion d'agent et système multi-agents

Il n'existe pas, actuellement, une définition de l'agent qui fasse foi dans le monde de l'intelligence artificielle distribuée. Il est donc nécessaire, pour avoir une bonne vision de ce concept, de confronter plusieurs de ces définitions. Nous allons exploiter trois d'entre elles.

Jacques Ferber [Ferber, 1995] définit un agent comme étant une entité physique

ou virtuelle évoluant dans un environnement dont il n'a qu'une représentation partielle et sur lequel il peut agir. Il est capable de communiquer avec d'autres agents et est doté d'un comportement autonome.

Cette définition aborde une notion essentielle : l'autonomie. En effet, ce concept est au centre de la problématique des agents. L'autonomie est la faculté d'avoir ou non le contrôle de son comportement sans l'intervention d'autres agents ou d'êtres humains. Une autre notion importante abordée par cette définition concerne la capacité d'un agent à communiquer avec d'autres.

Selon Yves Demazeau [Demazeau and Costa, 1996], un agent est une entité réelle ou virtuelle dont le comportement est autonome, évoluant dans un environnement qu'il est capable de percevoir et sur lequel il est capable d'agir, et d'interagir avec les autres agents.

Cette définition introduit l'interaction qui, comme nous le verrons par la suite, est le moteur des systèmes multi-agents. En effet, l'interaction suppose la présence d'agents capables de se rencontrer, de communiquer, de collaborer et d'agir.

Pour Mickael Wooldridge [Wooldridge, 1999], un agent est un système informatique capable d'agir de manière autonome et flexible dans un environnement.

Par flexibilité on entend :

– Réactivité : un système réactif maintient un lien constant avec son environnement et répond aux changements qui y surviennent.

– Pro-activité : un système pro-actif (aussi appelé téléonomique) génère et satisfait des buts. Son comportement n'est donc pas uniquement dirigé par des événements.

– Capacités sociales : un système social est capable d'interagir ou de coopérer avec d'autres systèmes.

Ces définitions sont le résultat de différentes approches. Il existe en fait plusieurs types d'agents, qui, selon les capacités qu'ils possèdent, seront qualifiés de réactifs, cognitifs ou hybrides.

Capacités cognitives. Les agents à capacités cognitives proviennent d'une métaphore du modèle humain. Ces agents possèdent une représentation explicite de

leur environnement, des autres agents et d'eux-mêmes. Ils sont aussi dotés de capacités de raisonnement et de planification ainsi que de communication. Ces agents sont structurés en société où il règne donc une véritable organisation sociale. Le travail le plus représentatif de cette famille d'agent porte sur le modèle BDI (Believe Desire Intention) [Rao and Georgeff, 1995]. Les sources de ces travaux sont les sciences humaines et sociales.

Capacités réactives. Les agents à capacités réactives ne possèdent pas de moyen de mémorisation et n'ont pas de représentation explicite de leur environnement : ils fonctionnent selon un modèle stimuli/réponse. En effet, dès qu'ils perçoivent une modification de leur environnement, ils répondent par une action programmée. L'exemple le plus célèbre est celui de la fourmilière étudié par Alexis Drogoul [Drogoul, 1993]. Les sources des travaux sur ce type d'agents sont les sciences de la nature et de la vie.

Modèle d'agent hybride. Les agents hybrides sont des agents ayant des capacités cognitives et réactives. Ils conjuguent en effet la rapidité de réponse des agents réactifs ainsi que les capacités de raisonnement des agents cognitifs. Cette famille regroupe donc des agents dont le modèle est un compromis autonomie/coopération et efficacité/complexité. Pour illustrer cette famille, nous pouvons citer l'architecture ASIC [Boissier and Demazeau, 1996] utilisée pour le traitement numérique d'images, l'architecture ARCO [Rodriguez, 1994] créée dans le cadre de la robotique collective et l'architecture ASTRO [Occello and Demazeau, 1998] développée pour être utilisée dans les systèmes multi-agents soumis à des contraintes temporelles.

Système multi-agents. Un système multi-agents est un ensemble d'agents qui évoluent dans un environnement commun. Dans [Weiss, 1999], Gerhard Weiss définit l'intelligence artificielle distribuée comme étant l'étude, la conception et la réalisation de systèmes multi-agents qu'il présente comme étant des systèmes dans lesquels des agents intelligents interagissent et poursuivent un ensemble de buts ou réalisent un ensemble d'actions.

2.3.2 Les agents d'informations

Un agent d'information est défini par Klusch comme le développement et l'utilisation efficace des entités informatiques autonomes, appelées les agents intelligents de l'information, qui ont accès aux sources d'informations multiples et hétérogènes, et géographiquement distribuées comme dans l'Internet ou l'Intranets de corporation. La tâche principale de tels agents est d'exécuter des recherches proactives afin de maintenir, négocier l'information appropriée au nom de leurs utilisateurs ou d'autres agents. Ceci inclut des qualifications telles que la recherche, l'analyse, la manipulation et la fusion de l'information hétérogène aussi bien que la visualisation et le guidage de l'utilisateur dans l'espace disponible et individuel de l'information. Ces agents fournissent également l'accès intelligent à une collection hétérogène de sources d'informations [Klusch, 2001].

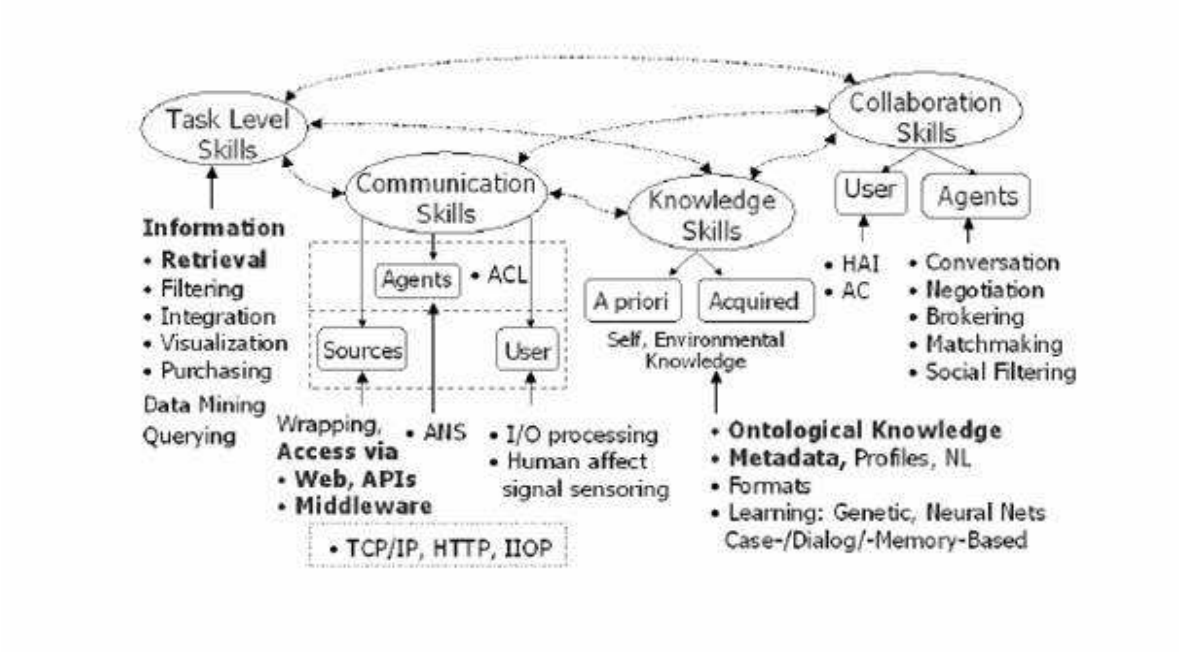


Figure 6 Compétence fondamentale des agents d'information

Les agents d'information peuvent être classés par catégorie dans différentes classes selon un ou plusieurs des dispositifs suivants [Klusch, 1999] ;

1. Les agents non coopératifs ou coopératifs de l'information, selon les capacités des agents à coopérer l'un avec l'autre pour l'exécution de leur tâche. Plusieurs protocoles et méthodes sont disponibles pour réaliser la coopération des agents

autonomes de l'information dans différents scénarios, comme la tâche de délégation hiérarchique et la négociation décentralisée.

2. Les agents adaptatifs d'information peuvent s'adapter aux changements des réseaux et des environnements de l'information. Les exemples d'un tel genre d'agents sont les assistant personnels apprenant sur le Web.

3. Les agents d'information raisonnables se comportent de manière utilitaire au sens économique. Ils agissent et peuvent même collaborer ensemble pour augmenter leurs propres avantages. Un domaine principal d'application d'un tel genre d'agents est le commerce électronique sur l'Internet.

4. Les agents mobiles de l'information peuvent voyager de façon autonome par l'Internet. De tels agents peuvent permettre la réduction du transfert de données parmi des serveurs de l'information, des applications.

2.3.3 Système multi-agents à la recherche d'information

Une grande partie des applications des Système Multi-agents est dans le domaine de la recherche d'information. Nous en présentons ici les systèmes les plus récents.

2.3.3.1 BONOM

Le système **BONOM** (Cazalens & Lamarre, 2001) (Cazalens et al., 2002) est un système multi-agents développé à l'université de Nantes pour la recherche d'informations et de connaissances distribuées sur le Web. Il propose :

- ✓ Une organisation d'agents selon une hiérarchie de domaines informationnels (thèmes). Les agents sont structurés en groupes relevant chacun d'un domaine. Le système dispose principalement de trois types d'agent: des agents sites, des agents intermédiaires et des agents utilisateurs ;
- ✓ Des mécanismes d'analyse de sites (situés sur les agents sites) qui permettent d'indexer le contenu d'un site Web en fonction d'ontologie(s) représentative(s) des domaines couverts par le site ;
- ✓ Des protocoles d'interaction permettant, à travers une recherche, d'atteindre les sites les plus pertinents. Il est à noter que ce sont les agents sites qui

s'inscrivent auprès d'agents succeptibles de leur envoyer les requêtes auxquelles ils peuvent répondre.

2.3.3.2 UMDL

C'est un système d'informations coopératif pour la recherche des documents dans une librairie digitale. Ce système (UMDL) [Weinstein *et al.*, 1999] est structuré comme une collections d'agents qui peut acheter et vendre des services à l'un ou l'autre (ou à l'utilisateur) en utilisant une infrastructure de commerce et de communications pour fournir les mécanismes par lesquels une bibliothèque numérique peut continuellement se modifier pendant que les utilisateurs, le contenu, et les services vont et viennent.

Dans le système UMDL, l'agent dédié à l'utilisateur demande à l'agent de Planification de l'aider pour constituer une équipe d'agents pour satisfaire une requête. L'agent de planification utilise trois autres agents pour l'aider ainsi. L'agent de Sujets fournit une hiérarchie des termes de plus en plus larges ou étroits qui sont employés par les agents de Collecte pour décrire les contenus de leurs collections. L'agent de Thesaurus aide la projection de la requête avec ces descriptions. Les descriptions sont enregistrées et associées aux adresses d'agent par l'agent Registry. Après avoir comparé la requête aux descriptions de collection, l'agent de Planification recommande un ou plusieurs agents de Collecte à l'agent de l'utilisateur, qui interagit alors directement avec des agents de Collecte pour traiter la requête. Notons que les agents poursuivent différents objectifs et sont capables de dire "non" à une demande. Par exemple, un agent pourrait refuser une demande s'il est très occupé, ou s'il n'est pas très bon pour satisfaire le genre particulier de demande.

Dans UMDL, la structure des messages et des conversations est implémentée selon KQML (Knowledge Query and Manipulation Language).

2.3.3.3 Calvin

Calvin [Bauer and Leake 2002] est un système multi-agents pour la recherche d'informations personnelles. Ce système observe des utilisateurs quand ils accèdent à des documents, trouve proactivement les documents relatifs, et fournit une

interface unifiée à l'environnement de l'information. Ce système se compose de deux parties :

- **Geneva** : la couche de « middleware» fournissant des services de base de SMA. Elle fournit la communication entre agents, l'authentification, et la fonctionnalité d'encryptage, et emploie des spécifications de « open XML » pour permettre de multiples types d'agents et un ensemble de tâches.
- **Calvin** : Les agents de Calvin implémentent individuellement des abstractions de haut niveau des parties de la tâche de recherche d'information, intégrant l'information des sources disparates.

Les agents dans le haut niveau sont divisés en deux groupes. Les agents d'analyse génèrent et diffusent des descriptions du contexte actuel de l'utilisateur. Les agents de recherche emploient ces descriptions de contexte pour questionner les moteurs de recherche standards. Les résultats sont envoyés de nouveau à l'agent de l'interface utilisateur qui les suggère à l'utilisateur. Concrètement :

Agents d'Interface :

L'agent *Calvin Web* fournit la communication bi-directionnelle entre le SMA et le navigateur Web d'un utilisateur. Les *Agents d'Analyse* analysent le comportement de l'utilisateur et développent un profil d'utilisateur reflétant les intérêts de l'utilisateur.

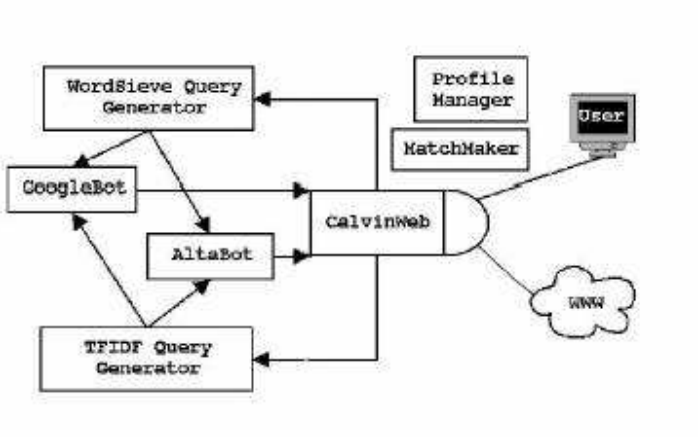


Figure 7 Architecture d'agent de Calvin

L'agent TFIDF applique l'algorithme d'analyse des textes « Term Frequency/Inverse Document Frequency » pour se renseigner sur les intérêts de

l'utilisateur et l'agent WordSieve applique l'algorithme « WordSieve » pour développer un profil d'utilisateur et se renseigner sur l'intérêt courant de l'utilisateur. L'agent DocStats accumule des informations statistiques sur les documents qu'un utilisateur a consulté.

Agents de recherche : Ces agents exécutent des recherches de fond pour l'utilisateur basé sur leurs profils d'utilisateur. *AltaBot* exécute des recherches de fond sur Alta Vista et *GoogleBot* exécute des recherches de fond sur Google.

2.3.3.4 NetSA

NetSA (Networked Software Agent) [Cote and Troudi, 1998] [Cote et al., 2001] est une architecture de système multi-agents pour la recherche d'information dans des sources hétérogènes et réparties. Cette architecture est composée de trois couches : la couche de communication avec l'utilisateur, la couche de traitement de l'information et la couche d'interrogation et d'extraction d'informations.

Ce système comporte plusieurs types d'agents comme :

- Un agent utilisateur en charge du recueil et du filtrage des informations provenant et allant vers l'usager;

-Un agent courtier (broker) servant de répertoire pour les agents qui évoluent au sein de NETSA;

- Des agents ressources reliés chacun à une ressource d'informations et pouvant rapatrier et mettre à jour les données;

- Un agent d'exécution en charge de la décomposition des tâches et du suivi du déroulement d'exécution des différentes sous-tâches;

- Un agent ontologie en charge du maintien de la cohérence des concepts utilisés par les agents.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les connaissances de base concernant notre domaine de recherche. Il s'agit des principes et des technologies du Web sémantique : l'architecture en couches du Web Sémantique, l'ontologie, les langages du Web Sémantique comme XML, RDF, OWL.

Nous avons focalisé notre étude de l'état de l'art sur les systèmes d'annotation manuelle, semi-automatique et automatique qui s'appuient sur de différentes techniques. Nous avons présenté les notions de base d'agents, de système multi-agents, et leurs applications pour la recherche d'information.

Partie II:

Vers le système OntoWatch

3 La veille au CSTB

3.1 L'organisation de la veille au CSTB

Tel que conçu en 1999, l'objectif du projet de veille au CSTB réside dans la mise en place d'une veille collective dont le but principal est de collecter et d'analyser les informations détenues en interne. La surveillance des informations externes (publiées sur Internet ou dans des revues scientifiques…), vient compléter cette démarche mais celle-ci s'appuie avant tout sur les informations grises (non publiées) et informelles (notamment obtenues par les réseaux de contacts personnels) car elles sont récentes, peu connues et non accessibles à tous.

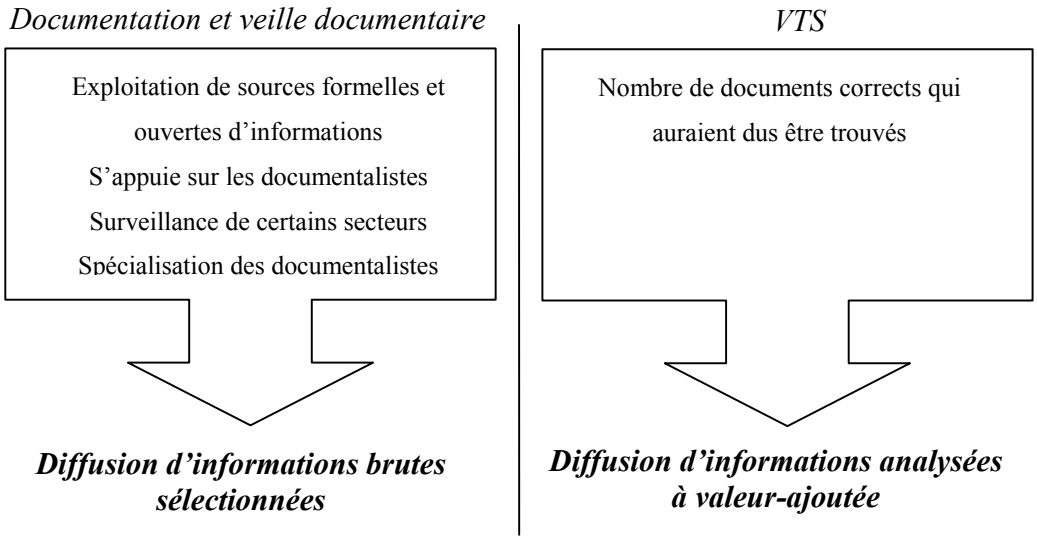


Figure 8 La veille documentaire et la veille technologique stratégique

Cette veille comprenait deux volets :

- la veille documentaire effectuée par les documentalistes, qui collectent,

traitent et diffusent les informations formelles,

- la veille technologique et stratégique (VTS) effectuée par les ingénieurs et chercheurs, qui vise notamment à collecter et qualifier les informations informelles qu'ils obtiennent au cours de leur vie professionnelle.

L'implication active des départements pour alimenter le système n'ayant pas eu les résultats escomptés, le travail de VTS se recentre aujourd'hui sur de la prestation offerte par les documentalistes du CSTB sur des projets de veille pour le compte des départements. La veille documentaire comprend :

- la veille technico-réglementaire,
- la veille normative (normes françaises et européennes),
- la veille technique & scientifique

3.2 Sources d'information concernées

La détermination du type et de la nature de ressources est très importante dans le processus de veille. Après avoir lu différents documents et consulté le site Intranet du CSTB, nous avons pu identifier les sources d'information suivantes :

- Bases de données
- Bases de données multidisciplinaires (par exemple : Science direct, PASCAL)
- Bases de données spécialisées (Avis techniques, CD-Reef)
- Bases de données brevets (Derwent, EspaceNet)
- Magazines professionnels
- Revues
    - généralistes (Moniteur des travaux publics & bâtiment, cahiers techniques du bâtiment,..)
    - spécialisées / scientifiques (Sols murs plafonds, Roofing Magazine)
- Sites webs
- Manifestations professionnelles
- Bulletins / Newsletters

3.3 Types de documents

Différents types de documents sont manipulés [eBib 2000] :

- **Articles** : scientifiques, de vulgarisation
- **Rapports** :

- rapports de recherche,
- rapports de consultance : réalisés pour le compte d'un client externe. En général, ce type de rapport reste confidentiel à cause des contraintes posées par le commanditaire.
- rapports d'activité, présentant le bilan annuel des activités (recherche, consultance, évaluation technique, etc.) des différents services du CSTB ou d'autres organismes extérieurs.
- rapports de stage.

- **Communications** : textes indépendants reprenant des interventions faites lors de congrès, séminaires, salons… Ces communications peuvent être rassemblées dans des actes publiés.
- **Thèses** : travaux de recherche visant à explorer, comprendre et synthétiser un problème ou un sujet et donnant lieu à un mémoire dont la soutenance devant un jury universitaire est nécessaire pour l'obtention du grade de docteur.
- Brevets
- Etudes de marché
- Normes
- Ouvrages bibliographiques
- Documents législatifs et réglementaires (loi, arrêté, etc.)

Les documents cités plus hauts peuvent être des documents en papier ou des documents électroniques dans divers formats : pages HTML (HyperText Markup Language), format PDF (Portable Document Format), format Microsoft Word et textes.

3.4 Processus de veille et le modèle de LESCA

La veille a pour principal objectif la valorisation de l'information. Elle fait passer l'information d'un état brut à un état élaboré par une série d'opérations

généralement intellectuelles : sélection de l'information à analyser, réunion de groupes de travail pour analyse, rédaction de compte-rendu, synthèses.

La veille technologique et documentaire chez CSTB est menée par des veilleurs qui sont des documentalistes, et en même temps des experts dans plupart de situations.

Dans cette section, nous proposons une description du processus de veille au CSTB (Centre Scientifique et Technique du Bâtiment) en reposant sur le modèle général de veille proposé par [Lesca, 2002]. Cette analyse permettra d'identifier les diverses phases où une ontologie ou des agents logiciels pourraient intervenir pour améliorer le processus de la veille technologique actuellement menée au CSTB.

Ciblage (Expression des besoins)

Le ciblage définit les thèmes sur lesquels la veille doit être déployée. Il s'agit de préciser les sujets qui doivent faire l'objet d'une surveillance active et de les hiérarchiser selon leur degré d'importance. L'activité de ciblage produit un profil de veille qui est un document descriptif élaboré en concertation avec le client qui permet de structurer les besoins exprimés en formalisant la thématique générale, les axes et types de veille l'intéressant.

Détection et Identification des sources d'information

Dans cette étape, le veilleur dresse une liste des sources à interroger. Il peut s'agir soit de sources déjà connues par le veilleur, soit de nouvelles sources trouvées grâce à des moteurs de recherche classiques qui effectuent en général une recherche par mots-clés.

Collecte, sélection, rapatriement des informations

Cette phase vise à rassembler des documents, des études sur les thèmes prédéfinis préalablement. Les documents sont téléchargés par des aspirateurs du Web ou via des navigateurs. L'étape de la collecte de l'information est une étape délicate, car les résultats de l'étape précédente risquent d'être probablement trop nombreux, entraînant la surabondance d'informations. La sélection des documents semblant pertinents joue un rôle important pour réduire la charge de l'étape d'analyse qui suivra immédiatement.

Analyse

Cette phase consiste en l'évaluation des informations collectées. Les acteurs concernés ici sont les documentalistes ainsi que les experts du domaine. Si les informations sont suffisantes pour répondre aux besoins, on peut passer à l'étape suivante.

Traitement et stockage

La lecture, l'extraction du contenu pertinent de faire une synthèse cohérente et porteuse de sens afin de rédiger des fiches d'information et constituer des « dossiers ressources ». Dans cette étape, soit les documentalistes soit les experts du domaine du CSTB peuvent créer des annotations sur les documents après la phase d'analyse. Les résultats de cette phase sont enregistrés afin d'être disponibles en cas de demande des utilisateurs.

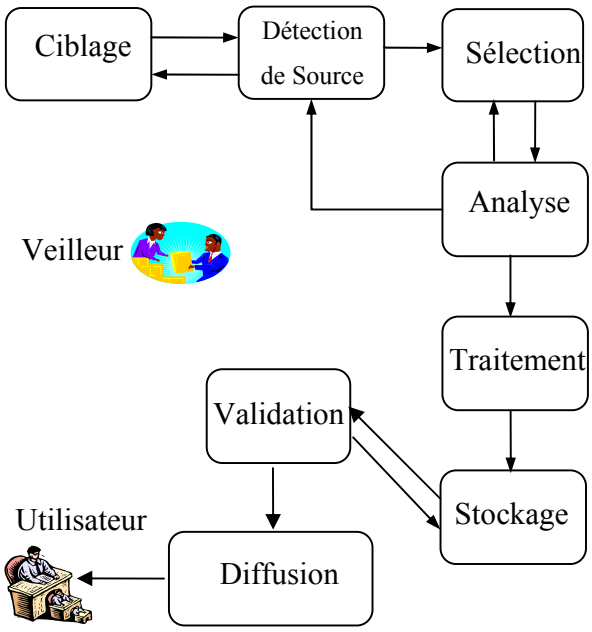


Figure 9 Le processus de veille au CSTB

Validation et Diffusion

Il est nécessaire de transmettre les résultats de la veille à l'utilisateur qui a demandé le service de veille. C'est seulement après sa validation qu'une diffusion générale à tous les utilisateurs de la veille peut avoir lieu. Pour la validation, les observateurs peuvent envoyer les résultats de la veille à certains experts du domaine.

3.5 Résultats de la veille

Les livrables de la veille peuvent être :

- des fiches d'information (stockées dans une base de données ou diffusées par mail),
- des synthèses,
- des dossiers ressources, des états de l'art,
- des bulletins d'information thématiques.

3.6 Outils et moyens techniques employés

La masse d'informations disponibles sur l'Internet est telle que seuls les outils automatisés de recherche et d'extraction d'information peuvent répondre aux contraintes de temps par une activité de veille. Il soulage et décharge les utilisateurs des tâches longues et répétitives.

Les trois outils les plus utilisés par les veilleurs pendant les tâches de veille sont :

- Des annuaires et des bookmarks,
- Des moteurs de recherches et des méta moteurs : Yahoo, Google, Kartoo, etc.,
- Des outils de analyse du documents.

3.7 Évolutions souhaitées du système de veille

Le système de veille actuel fonctionne fondamentalement en se basant sur les compétences et les expériences des acteurs de la veille, notamment du documentaliste. Dans presque toutes les étapes dans le processus de veille, le travail est réalisé manuellement, avec l'aide de quelques outils, dépendant de chaque veilleur. On ne peut bien sûr pas remplacer le rôle des experts humains dans l'analyse et le traitement des documents, qui sont actuellement trop difficiles à comprendre pour les machines, mais il faut chercher à améliorer le système sous plusieurs aspects. Concrètement, le CSTB a besoin d'un système qui :

- Aide à la modélisation et à l'automatisation des étapes de la recherche qui reposent sur : la connaissance tacite que le veilleur possède et son contexte de travail, les questions que se pose le veilleur pour pouvoir diriger sa recherche dans la bonne direction (objectifs de la recherche, usage fait des

informations….) et les conclusions qu'il en tire (type de documents recherchés…).

- Aide à l'automatisation de certaines tâches dans l'analyse et le traitement des documents trouvés, par exemple la tâche d'annotation (pour créer des méta-données du document).
- Permettre une meilleure gestion des documents internes et externes du CSTB, grâce à la recherche sémantique sur des bases d'annotation.
- Permettre d'automatiser la surveillance de certaines sources d'information

4 L'ontologie pour la veille

Comme nous l'avons expliqué dans le chapitre précédent, l'ontologie est le cœur de notre système de veille, autrement dit la veille dans le nouveau système est guidée par une ontologie. L'utilisation de l'ontologie dans certaines étapes du processus de veille permet d'améliorer le résultat de chaque étape et évidemment la performance globale du système. Nous pourrons constater que tous les changements positifs du système concernent plus ou moins l'ontologie.

La qualité de l'ontologie va sans doute influencer les résultats de plusieurs tâches ainsi que l'efficacité du système. Assurer la qualité et la disponibilité des vocabulaires dans l'ontologie est un critère important dans le but de développer un système de veille efficace. Dans ce chapitre nous présentons les travaux réalisés pour construire une ontologie dédiée à notre système de veille. Il se décompose en 3 sections : dans la première section nous allons introduire la structure de l'ontologie à construire. Puis dans le deuxième section nous allons aborder la réutilisation d'une partie de l'ontologie O'CoMMA dans la nouvelle ontologie O'Watch. Ensuite nous présenterons nos travaux pour enrichir les vocabulaires en transformant les termes dans des thésaurus de domaines en des concepts et des propriétés.

4.1 Démarche

La construction de l'ontologie est encore un sujet de recherche qui intéresse de nombreux chercheurs dans le domaine d'intelligence artificielle depuis plusieurs années. Ce n'est évidemment pas une tâche simple et totalement automatique. Dans le cadre de cette thèse, nous ne nous sommes pas focalisé sur la recherche d'une nouvelle méthodologie. Nous nous sommes particulièrement inspirés des méthodologies proposées par Gomez [Gomez-Perez, 1998], Fernandez et [Fernandez et al., 1997] et Uschold [Uschold et al., 1995]. Le but final est de disposer d'une « bonne» ontologie qui fournit les primitives sémantiques du système. Nous n'avons pas commencé ce travail à partir de zéro, et avons hérité des résultats des travaux précédents réalisés au sein de l'équipe Acacia concernant la construction de l'ontologie.

Nous présentons ici les tâches réalisées pour obtenir l'ontologie de veille O'Watch :

- Identifier le but et le domaine couvert de l'ontologie. Déterminer la structure et les composants principaux ;
- Réutiliser des ontologies existantes;
- Déterminer les terminologies réutilisables;
- Collecter les sources d'informations disponibles pour construire l'ontologie : documents, experts, thésaurus;
- Compléter ces ontologies par le vocabulaire dans ces sources ;
- Représenter les vocabulaires en RDFS;
- Utiliser et tester l'ontologie et puis la corriger en cas de besoin.

4.2 Analyse du contexte et identification des parties principales

Le but de l'ontologie est de fournir des vocabulaires sémantiques explicites satisfaisant les besoins du système de veille. Pour arriver à ce but, il est important de déterminer tout ce qui dans l'ontologie est nécessaire et utile pour le système dans chaque étape du processus de la veille technologique au CSTB. Nous pouvons

obtenir la réponse à cette question à partir de l'analyse des rôles de l'ontologie dans ces étapes, comme nous l'avons présenté dans les chapitres précédents. En résumé, le système a besoin d'une ontologie capable de fournir les vocabulaires :

- pour préciser la formulation de la requête du veilleur, lever l'ambiguïté du contexte de recherche et explicitement décrire les sources d'information, les profiles de veille,
- rattacher les documents trouvés à des thèmes spécifiques,
- pouvoir bien annoter le document et également le profil des participants d'une veille.

Nous estimons intuitivement que l'ontologie pour la veille doit comprendre :

- une partie dédiée aux domaines de la veille tels que la construction, des bâtiments et des équipements, etc. De toute manière on ne peut pas assurer de disposer d'une ontologie qui couvre complètement la connaissance sur un domaine, surtout quand on fait de la veille sur le Web qui est en constante évolution. Cette partie doit être mise à jour avec l'évolution des terminologies dans les domaines à surveiller.
- une partie dédiée aux tâches spécifiques de la veille, et concernant tous les concepts et propriétés nécessaires pour la description des aspects importants de ces tâches.

4.3 Réutilisation des ontologies

La réutilisation des ontologies est évidemment attractive mais difficile en même temps. Elle réduit le temps et les travaux pour le créateur de l'ontologie mais elle exige la cohérence au niveau de la conceptualisation entre l'ontologie désirée et l'ontologie réutilisée. Chaque ontologie est dédiée à un objectif particulier et l'importation automatique des vocabulaires est impossible. La réutilisation dirigée par un humain est un choix plus raisonnable. Pour plusieurs raisons qui seront expliquées plus tard, nous avons réutilisé l'ontologie O'CoMMA comme une partie importante dans l'ontologie souhaitée.

4.3.1 Ontologie O'CoMMA

L'ontologie O'CoMMA a été développée dans le projet européen CoMMA (Corporate Memory Management through Agents) [Gandon, 2002]. L'objectif principal de ce projet était de mettre en application et de tester un système de gestion de mémoire d'entreprise basé sur les systèmes multi-agents (SMA), et d'aborder en particulier les questions suivantes :

- Faciliter l'insertion de nouveaux employés dans l'entreprise.
- Assister le processus de veille technologique.

L'ontologie a été conçue et utilisée pour annoter sémantiquement les ressources des intrawebs d'une entreprise dans le domaine des Télécommunications et d'un organisme de recherche sur le Bâtiment (CSTB). O'CoMMA est formalisée dans le langage RDF Schema et contient : 470 types de concepts organisés en une hiérarchie d'une profondeur maximale de 13 liens de subsomption, 79 types de relations formant une hiérarchie d'une profondeur maximale de 2 liens de subsomption. O'CoMMA utilise 715 termes anglais et 699 termes français pour étiqueter ces primitives.

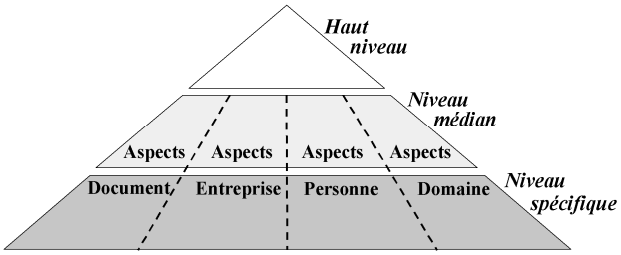


Figure 10 La structure de O'CoMMA

La figure 10 montre les trois couches structurant l'ontologie :

- une couche supérieure très générale qui ressemble aux autres ontologies de haut niveau ;
- une couche médiane assez importante, divisée en une partie générique concernant le domaine des mémoires organisationnelles (documents, organisations, personnes...) et une autre consacrée au domaine d'application (ex. pour les télécommunications: WAP, mobile, etc.) ;

- une couche d'extensions spécifiques au scénario et à l'entreprise, avec des concepts complexes (ex. rapport d'analyse de tendance, carte d'itinéraire du nouvel employé).

4.3.2 Réutilisation de l'ontologie O'CoMMA

Le partage dans le but et le domaine appliqué de cette ontologie avec l'ontologie désirée est une raison importante pour sa réutilisation. Cette ontologie a été développée et évaluée avec l'aide des utilisateurs du CSTB. Une grande partie des vocabulaires est donc réutilisable. Après avoir analysé cette ontologie et étudié le besoin pour le nouveau système, nous avons décidé de garder la structure de base de cette ontologie, et de n'enlever que certaines petites parties qui apparaissaient comme moins importantes. La couche supérieure est gardée car elle contient des concepts communs et très abstraits, qui sont réutilisables dans le scénario d'application concernant la veille technologique et scientifique pour l'entreprise.
La couche médiane traite le domaine d'application. Elle est donc réutilisable pour le scénario de veille. La veille technologique concerne tous les domaines d'application. Même ceux qui n'intéressent pas actuellement le CSTB gardent encore leur valeur pour l'évaluation des algorithmes utilisant l'ontologie pour la recherche et l'annotation des documents.
La dernière couche contenant des concepts spécifiques n'est plus réutilisable dès que l'organisation, le scénario ou le domaine d'application change. Nous n'avons gardé que les concepts et propriétés concernant le scénario « aide à la veille technologique ».
La structure divisée en couches selon le niveau de spécificité de l'ontologie OCoMMA nous permet d'y intégrer des vocabulaires nécessaires provenant de plusieurs sources de données pour obtenir l'ontologie finale.

4.4 Enrichir l'ontologie O'CoMMA

Après avoir analysé O'CoMMA et le besoin des vocabulaires dans les étapes du processus de veille, nous nous sommes aperçus que la partie réutilisable de O'CoMMA n'est pas suffisante. Deux grandes parties de l'ontologie devaient être enrichies et alimentées avec de nouveaux concepts et propriétés:

a) La partie dédiée à la tâche de veille dans O'CoMMA, qui donne une spécification explicite sur les acteurs, les actions de veille et, plus important, les types de sources d'information et certains types de documents manquant dans O'CoMMA.
b) La partie dédiée au domaine d'application, autrement dit, domaine sur lequel la veille sera effectuée. A côté des domaines déjà existants dans O'CoMMA comme la télécommunication, la construction et le bâtiment (en général), les nouveaux domaines intéressants par les veilleurs du CSTB sont divers : "sécurité et incendie", "second oeuvre", "réutilisation de l'eau", etc.

4.4.1 Enrichir l'ontologie dédié à la tâche de veille

Nous avons suivi les étapes suivantes pour ajouter des nouveaux vocabulaires à la partie de l'ontologie O'CoMMA dédiée à la tâche de veille :

- Déterminer les groupes de types des concepts et des propriétés concernant la tâche de veille. Ils se composent des concepts et des propriétés décrivant : (i) la source d'information d'où le document est issu, (ii) le type du document à chercher, (iii) les rôles des acteurs participant à la veille et leur action..
- Collecter des documents et données : Grâce à l'aide des documentalistes et des veilleurs du CSTB, nous avons obtenu des documents décrivant la tâche de veille actuellement menée au CSTB. Les informations collectées et synthétisées à partir des documents nous ont fourni des termes très proches des concepts de l'ontologie.
- Conceptualiser ces termes, déterminer les relations de subsomption ainsi que les autres relations et les comparer avec les concepts et propriétés de la partie correspondante dans O'CoMMA pour trouver l'endroit (le concept parent) le plus approprié pour intégrer ces nouveaux vocabulaires. En fait, la terminologie décrivant les types de documents, les types des sources d'information, les acteurs était proche des labels correspondant aux concepts. Cette tâche a donc été relativement simple et directe.

La figure 11 montre quelques exemples des concepts dédiés aux types de documents dans la nouvelle ontologie.

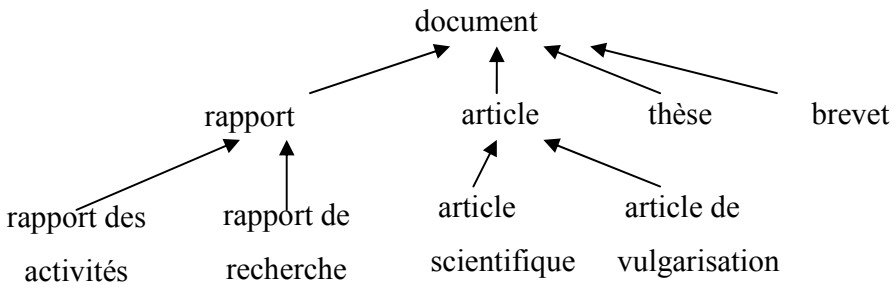


Figure 11 Concepts correspondant aux types de document

4.4.2 Enrichir l'ontologie dédiée aux domaines de veille

Les terminologies nécessaires pour décrire les domaines de veille ont été collectés à partir de deux sources principales : les documents internes du domaine du CSTB et les thésaurus du CSTB.

Pour travailler avec les documents internes, nous avons suivi les méthodes traditionnelles, reposant sur l'analyse de document soit manuellement, soit avec l'aide d'outils de traitement du langage naturel pour obtenir une liste des candidats terminologiques. La phase de conceptualisation et d'intégration se déroule normalement comme nous l'avons fait avec la partie dédiée à la tâche de veille.

La tâche de transformation des terminologies aux concepts et de détermination des relations entre eux est beaucoup plus difficile. Dans la section suivante, nous allons aborder le travail d'exploitation des vocabulaires dans un thésaurus pour construire une ontologie.

4.4.3 Transformation des vocabulaires de thésaurus en une ontologie

4.4.3.1 Qu'est ce que c'est un thésaurus ?

Définition :

Selon la définition de [ISO 2788 :1986] : *un thésaurus est le vocabulaire d'un langage d'indexation contrôlé, organisé formellement de façon à expliciter les relations à priori entre les notions (par exemple relations générique-spécifique)*.

Un thesaurus est une sorte de dictionnaire hiérarchisé ; un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. Il ne fournit qu'accessoirement des définitions, les relations des termes et leur choix l'emportant sur leurs significations. Un thesaurus est aussi un ensemble structuré de termes choisis pour leur capacité à faciliter la description d'un domaine et à harmoniser la communication et le traitement de l'information à son sujet.

Chaque terme appelé descripteur est aussi peu ambigu que possible et est préféré à des termes voisins (quasi-synonymie) ou synonymes, les non-descripteurs, pour tous les échanges significatifs.

Le thesaurus est traditionnellement un outil d'indexation et de recherche documentaire. Il traduit en langage normalisé tout concept décrivant le contenu d'un document. Il permet par la suite de retrouver ce document dans un ensemble d'informations décrit par ce vocabulaire contrôlé.

L'utilisation ou exploration d'un thesaurus peut se faire à l'aide de plusieurs modes de présentation :

- Liste(s) alphabétique(s) des termes ; pour une approche globale ou la recherche d'un terme particulier ;
- Liste(s) hiérarchique(s) des termes ; pour l'approfondissement d'une notion ;
- Liste(s) d'occurrences (liste permutée) ; pour la vérification de la pertinence d'un élément d'une expression utilisé comme descripteur.

4.4.3.2 Les relations

Les termes d'un thesaurus sont organisés hiérarchiquement. Tout thesaurus comporte au moins trois catégories de termes : les termes génériques et les termes spécifiques qui doivent être utilisés comme descripteurs ; les termes équivalents qui sont considérés comme non-descripteurs selon les conventions du thesaurus.

- Les termes génériques sont repérés généralement par le sigle TG ; ils désignent les entités ou concepts principaux en référence aux autres termes et au domaine considéré ;

- Les termes spécifiques sont repérés généralement par le sigle TS ; ils précisent et identifient les entités ou concepts particuliers à l'intérieur du champ sémantique d'un terme générique donné ;
- Les termes équivalents sont repérés généralement par le sigle EP comme abréviation de Employé Pour ; ce sont des variantes des termes spécifiques (synonymie ou quasi-synonymie). Le terme à préférer au terme Employé Pour est indiqué par le symbole EM ou EMP comme abréviation de Employer.

Les relations entre les termes sont de trois types : relation hiérarchique, relation d'équivalence, relation d'association.

La relation d'équivalence

Un même terme peut référer à plusieurs concepts et un même concept peut être désigné par plusieurs termes. La relation d'équivalence suppose le renvoi au descripteur pour un terme non préférentiel.

Les relations hiérarchiques

Les relations hiérarchiques constituent la structure essentielle du thésaurus autour de laquelle s'organise le classement des informations. Les relations hiérarchiques servent à exprimer les rapports de généralisation d'un domaine sémantique plus large et de spécialisation d'un domaine sémantique plus restreint, entre les concepts représentés par les descripteurs.

La relation d'association

Un concept est associé à un autre comme étant relatif au même sujet, et pour en suggérer un autre aspect. La relation d'association est une relation symétrique entre deux descripteurs désignant des concepts qui, bien que non liés entre eux par une équivalence linguistique ou une hiérarchie, sont susceptibles de s'évoquer mutuellement, par association d'idées.

4.4.3.3 Du thésaurus à l'ontologie

Il y a parfois une confusion entre les deux notions : thésaurus et ontologie. D'un point de vue rigoureux un thésaurus n'est pas vraiment une ontologie. Concrètement, un thésaurus relie des termes entre eux selon des relations précises : synonyme, homonyme, hiérarchie, terme associé. L'ontologie ajoute des règles et

des outils de comparaison sur et entre les termes, groupes de concepts et relations : équivalence, symétrie, contraire, cardinalité, transitivité... Ainsi, l'ontologie est une étape supérieure au thésaurus selon «l'ontology spectrum ».

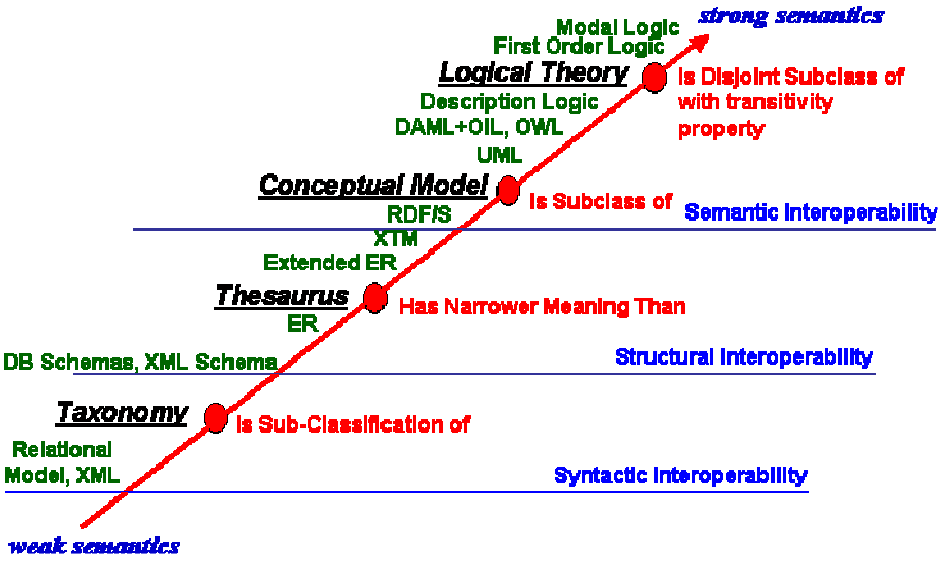


Figure 12 Thésaurus et Ontologie dans le spectre d'ontologie

Le « spectre d'ontologie » décrit un continuum des modèles sémantiques d'expressivité et de complexité croissantes : taxonomie, thesaurus, modèle conceptuel, et théorie logique. Dans ce spectre d'ontologie, un thesaurus est un ensemble des termes reliés, où les relations entre les termes dans la hiérarchie de thesaurus sont interprétées en tant que relations « narrower – broader ». Les ontologies ajoutent plus d'expressivité dans la spécification des rapports entre les concepts.

La différence entre ontologie et thésaurus s'exprime dans leurs définitions formelles :

Définition 1. *Une ontologie est un triplet O = (C, R, isa) défini comme suit :*

1. *$C=\{c_1, c_2, ..., c_n\}$ est un ensemble de concepts. Chaque concept se rapporte à un ensemble d'objets dans le monde réel (instances)*
2. *$R=\{r_1, r_2, ..., r_m\}$ est l'ensemble des rôles typés binaires entre des concepts*
3. *isa est un ensemble des relations de héritage entre des concepts.*

Définition 2 . Un thésaurus est un couple T = (D, rt) tel que

1. *D= {t1, t2, ..., tn } est un ensemble de termes*
2. *rt est une relation binaire entre les termes de 5 genres :*
   - *généralisation*
   - *instance*
   - *partitive*
   - *équivalence*
   - *associative*

La relation hiérarchique entre les termes dans le thésaurus inclut trois premiers types dans la définition (généralisation, instance, partitive) où seulement la première est la relation de subsomption. Alors, bien que ces thesaurus soient hiérarchiques, il ne s'agit pas vraiment d'ontologies puisque le lien hiérarchique dans les thesaurus pouvait être interprété comme "si nous recherchons un document parlant de $sujet_1$, un document parlant de $sujet_2$ est utile".

Par exemple :

TG terme générique

TS terme spécifique

Exemple :

Champs sémantique ***EQUIPEMENTS SANITAIRES***

TG niveau 1	INSTALLATION SANITAIRE
TS niveau 1	EQUIPEMENT DE PLOMBERIE
TS niveau 2	PLOMBERIE
TS niveau 3	CLAPET
TS niveau 3	JAUGE DE NIVEAU

Mais cela ne signifie pas que $sujet_2$ est nécessairement un sous-concept de $sujet_1$. Ce lien hiérarchique n'est pas toujours un lien de subsomption. Par exemple, dans le thesaurus "sécurité et flamme", un lien hiérarchique relie directement le sujet "zone de flamme" à "émissivité de flamme", à "taille de flamme" et à "température de flamme" : ce lien hiérarchique n'est clairement pas un lien de subsomption.

Pour cette raison la traduction directe et automatique des termes dans un thésaurus en des concepts dans une ontologie est impossible et la transformation manuelle

devient inévitable. Nous présentons ici les étapes effectuées pour intégrer des vocabulaires à partir d'un thésaurus dans une ontologie existante.

Avec chaque terme descripteur T dans le thésaurus nécessaire à ajouter à l'ontologie :

- Déterminer la position dans l'ontologie pour le nouveau concept T (le concept parent)
- Ajouter T dans l'ontologie comme concept T
- Examiner toutes les relations de descripteur T avec les autres termes {T'} dans le thésaurus
- Si la relation entre (T, T') est la relation d'équivalence : T' est ajouté dans l'ontologie comme un étiquette de concept T. Il joue le rôle d'un synonyme pour la formulation des requêtes.
- Si la relation entre (T, T') est la relation d'association : T' est ajouté dans l'ontologie comme un concept. On ajoute à l'ontologie une relation «RelatedTo» entre T et T'
- Si la relation entre (T, T') est la relation hiérarchique :
    - Déterminer si la relation entre ces termes est exactement la subsomption « est-un » (« is a »)
    - Ajouter T' à l'ontologie comme le concept enfant (plus spécialisé) de T

Continuer les mêmes étapes pour le concept T'.

Pour les relations « part-of » entre les termes dans le thésaurus, nous avons décidé de ne pas les intégrer dans l'ontologie car ces relations ne seront pas exploitées par les algorithmes pour améliorer le système de veille.

Par exemple pour les termes dans le thésaurus de « gros œuvre et second œuvre »

TS Niveau 2	EQUIPEMENT SANITAIRE
TS Niveau 3	BAIGNOIRE
TS Niveau 3	BIDET
TS Niveau 3	DOUCHE
TS Niveau 3	LAVABO

sont ajoutés dans l'ontologie O'CoMMA comme suit :

```
<rdfs:Class rdf:ID="SanitaryEquipmentTopic">
 <rdfs:subClassOf rdf:resource="#EquipmentTopic"/>
```

```
 <rdfs:label xml:lang="en">sanitary equipment</rdfs:label>
 <rdfs:label xml:lang="fr">équipement sanitaire</rdfs:label>
</rdfs:Class>
<rdfs:Class rdf:ID="LavatoryBasinTopic">
 <rdfs:subClassOf rdf:resource="#SanitaryEquipmentTopic"/>
 <rdfs:label xml:lang="en">lavatory basin</rdfs:label>
 <rdfs:label xml:lang="fr">lavabo</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="BidetTopic">
 <rdfs:subClassOf rdf:resource="#SanitaryEquipmentTopic"/>
 <rdfs:label xml:lang="en">bidet</rdfs:label>
 <rdfs:label xml:lang="fr">bidet</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="ShowerTopic">
 <rdfs:subClassOf rdf:resource="#SanitaryEquipmentTopic"/>
 <rdfs:label xml:lang="en">shower</rdfs:label>
 <rdfs:label xml:lang="fr">douche</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="BathTubTopic">
 <rdfs:subClassOf rdf:resource="#SanitaryEquipmentTopic"/>
 <rdfs:label xml:lang="en">bathtub</rdfs:label>
 <rdfs:label xml:lang="fr">baignoire</rdfs:label>
</rdfs:Class>
```

4.5 L'ontologie O'Watch

L'ontologie O'Watch est le résultat obtenu par les travaux présentés dans les sections précédentes. Comme O'CoMMA elle est formalisée en RDFS, et comprend 586 concepts et 86 propriétés. Les nouveaux vocabulaires ajoutés à l'ontologie sont extraits des plusieurs sources d'information :

- Les documents intérieurs du CSTB (les rapports sur la tâche de veille, les informations fournies par le service de documentation, les document contenant des vocabulaires d'un domaine spécifique…)
- Les thésaurus du CSTB : thésaurus « gros œuvre et second œuvre », thésaurus « sécurité et incendie » et « recyclage de l'eau »
- Le thésaurus Canadien de la science de construction et Bâtiment.

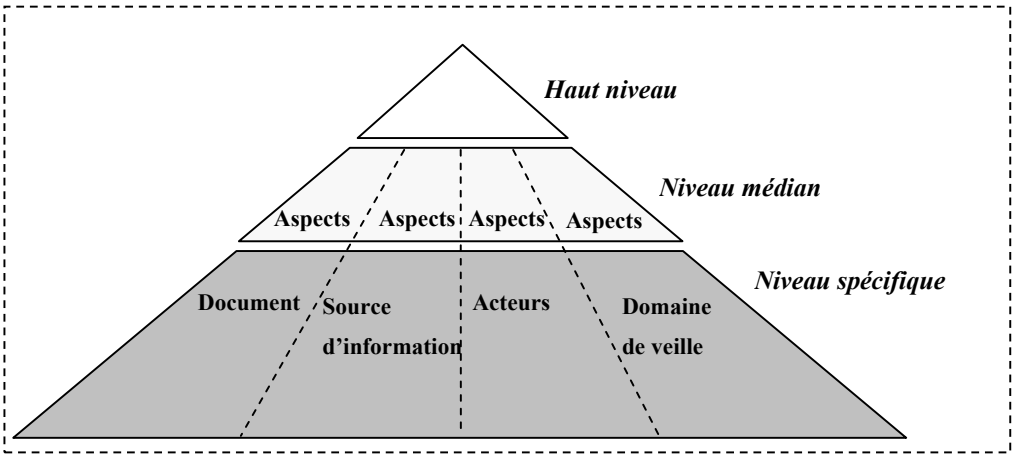


Figure 13 La structure de l'ontologie Watch.

D'une manière générale, l'ontologie O'Watch garde la structure de l'ontologie O'CoMMA. Elle se divise en trois couches : supérieure, médiane et extensions. La couche supérieure reste presque inchangée car elle comprend des concepts très abstraits, mais nécessaires pour définir des autres concepts plus spécifiques. La plupart des changements ont été effectué dans les deux couches qui restent. Les vocabulaires dans O'CoMMA qui ne concernent pas le scénario de veille technologique sont enlevés et à l'inverse les nouveaux concepts nécessaires sont ajoutés. Contrairement à O'CoMMA, l'ontologie O'Watch, comme le but défini, est une ontologie utile aux travaux de veille. Les domaines couvert sont donc plus concrets et spécifiques : Acteurs de veille, Source d'information, Document et les thèmes (domaine) de veille. Nous pouvons apercevoir une correspondance entre ces quatre parties avec celles de l'ontologie O'CoMMA : Personne, Document, Entreprise, Domaine.

Précisément, la partie *Domaine de veille* est le résultat de la réutilisation des vocabulaires dans la partie Domaine de O'CoMMA et l'intégration des terminologies dans les nouveaux domaines de veille. La partie Acteurs hérite d'une partie du vocabulaire de la partie Personne et est enrichie avec des nouveaux concepts concernant les acteurs de veille. Une partie complètement nouvelle est dédiée à la description : Source d'information, elle est utile non seulement pour

adapter les agents à la source sur laquelle ils fonctionneront (si l'on dédie un agent à une source donnée ou à plusieurs sources), mais également pour annoter les sources déjà connues par le système. Les annotations sur une source d'informations pourront fournir des informations supplémentaires et utiles quand l'on annotera les documents qui appartiennent à ces sources. La partie Document complète les concepts et ontologies dans O'CoMMA avec des nouveaux concepts désignant les types de documents relié à la veille. Les tables suivantes illustrent des exemples de vocabulaires dans les différentes parties de l'ontologie Watch.

Partie d'ontologie O'Watch dédiée aux documents

document
abstract
article
article scientifique
article de vulgarisation
avis technique
brève
brevet
commnication
norme
ouvrage bibliographique
rapport
rapport de recherche
rapport de consultance
raport de stage
rapport d'activité
proceeding
thèse
document législatif
page web
…. à voir dans l'ontologie pour plus de concepts

Partie dédiée à la source d'information

source d'information
base de données
base de données multidisciplinaire
base de données spécialisée
base de donnée de brevets

manifestation professionnelle
bibliothèque électronique
révues généralistes
révues spécialisé
site web
…..

La partie dédieé aux domaines différents de veille

materiaux de construction			
	beton		
		beton à aire occlus	
		beton hydrocarbone	
		beton cellulaire	
		beton gaz	
		beton mousse	
		beton coloré	
	métal		
		métal ferreux	
		métal non ferreux	
		fonte	
	mortier		
	papier		
	pièrre		
	polimer		
	….		

analyse et traitement de l'eau				
	traitement de l'eau			
		traitement de l'eau potable		
		purification de l'eau		
		traitement des eaux usées		
		assainissement des eaux		
			assainissement collectif des eaux	
			assainissement semi-collectif des eaux	
			assainissement autonome des eaux	
		désinfection de l'eau		
		traitement tertiaire		
		traitement autonome de l'eau		
	….			

equipement		
	equipement sanitaire	
		baignoire
		bidet

	douche	
	lavabo	
chauffage		
	chauffage de l'eau	
	chauffage solaire	
	aerothème	
		aerothème centrifuge
		aerothème à projection
….		

4.6 Conclusion

La construction de l'ontologie est une tâche qui exige l'expérience et l'expertise des ingénieurs des connaissances. Bien que de nombreux travaux récents démontrent l'effort entrepris pour automatiser certaines étapes de construction, le rôle des acteurs humains reste encore décisif pour construire une ontologie de qualité. Dans ce chapitre, nous avons décrit notre expérience de développement d'une ontologie dédiée à la veille technologique. Les étapes sont un peu différentes par rapport aux méthodes traditionnelles. A partir de l'analyse des rôles de l'ontologie à construire dans le système de veille, nous avons déterminé les parties de vocabulaires indispensables, puis en analysant une ontologie existante réutilisable, nous avons obtenu la structure de l'ontologie O'Watch. La majorité des nouveaux vocabulaires intégrés dans O'Watch proviennent des thésaurus, qui ont une sémantique proche de l'ontologie mais pas identique. La transformation automatique d'un thésaurus vers une ontologie est faisable seulement quand toutes les relations hiérarchiques dans le thésaurus sont des relations de subsomption, ce qui n'est pas le cas des thésaurus du CSTB. La distinction entre une relation hiérarchique « partie-de » et une relation de subsomption dans le thésaurus demande l'intervention de l'expert humain. L'étape d'évaluation et de validation d'une ressource terminologique ne peut se concevoir, nous l'avons dit, que par l'usage. La validation de notre ontologie doit donc se faire au regard des tâches qui lui sont attribuées, principalement l'aide à la reformulation de requête, et à l'annotation des résultats de la recherche d'information.

5 Architecture du système de veille OntoWatch

La raison d'être d'un système d'information est de recueillir, traiter, mettre à jour les informations issues de différentes sources et rendre accessibles ces informations aux utilisateurs. Un système d'information d'aide à la veille technologique et scientifique est destiné à la surveillance de l'environnement de l'entreprise, permettant de maîtriser la collecte et le traitement des informations de type scientifique et technique.

Si la recherche documentaire sur le Web à l'aide de mots clefs offre des performances intéressantes en matière de rapidité de traitement, elle a des faiblesses inhérentes : la précision (la capacité du système à ne trouver que les documents pertinents) et le rappel (la capacité du système à trouver tous les documents pertinents).

Le veilleur est souvent confronté au fait que : tantôt certains documents sont renvoyés en réponse de manière erronée, tantôt des documents pertinents manquent dans les documents recueillis. De plus, ces résultats ne sont pas directement exploitables et nécessitent un travail considérable d'analyse des documents sélectionnés pour extraire l'information pertinente. La prise en compte de l'aspect sémantique fait espérer une solution efficace pour résoudre ce problème. Nos travaux se concentrent sur l'amélioration de deux grandes tâches dans le processus de veille : la recherche d'information et l'annotations des documents.

Nous présentons dans ce chapitre notre approche pour construire un système d'aide à la veille technologique appelé OntoWatch. Cette approche repose largement sur des technologies du Web Sémantique et sur la notion d'ontologie. Nous présentons ensuite les caractéristiques visées pour notre système dédié à la veille et enfin nous détaillons son architecture.

5.1 Rôles de l'ontologie pour améliorer le système de veille

En analysant les étapes dans le processus de veille, nous avons déterminé celles où l'ontologie pourrait intervenir afin d'améliorer la tâche de veille. L'exploitation d'une ontologie pourrait être utile dans plusieurs phases :

- Dans la *phase de ciblage :* L'utilisation des concepts dans l'ontologie peut aider à éviter l'ambiguïté du contexte de recherche, en permettant de préciser la formulation de la requête. Une telle ambiguïté peut être liée aux phénomènes linguistiques comme la synonymie ou l'homonymie. Avec une ontologie décrivant les diverses sources d'information, le système pourra lancer les agents spécifiques dédiés à chaque source, pour rechercher l'information.
- Dans la *phase de collecte* : L'ontologie peut être utile pour enrichir les requêtes d'utilisateurs, permettant de trouver plus de documents pertinents et de réduire le silence.
- Dans la *phase de traitement* : En dehors de certaines activités essentielles dans lesquelles les machines peuvent difficilement se substituer aux humains (par exemple faire un résumé ou une synthèse du contenu de document, etc..), les ontologies nous semblent pouvoir jouer un rôle potentiel dans la tâche d'annotation de document. Nous nous sommes focalisés sur cette direction.
- Dans la *phase de diffusion* : Selon le centre d'intérêt de l'utilisateur, décrit en utilisant l'ontologie, le système pourra automatiquement envoyer à un utilisateur donné des suggestions pour lire certains documents lui semblant pertinents .

Nous avons constaté que la performance des deux sous-tâches importantes pourrait être améliorée avec le support d'un système d'information. Il s'agit de la recherche des documents et de l'annotation des documents.

Pour pouvoir exploiter tous ces avantages, notre système dédié à la veille OntoWatch repose sur une ontologie et il est inspiré des technologies du Web sémantique. Tout d'abord, la sortie du système consiste en des annotations sémantiques (en langage RDF) sur les documents. Avec les annotations sémantiques RDF, au lieu de télécharger et de stocker les documents intéressants trouvés, le

système aura juste besoin de gérer leurs annotations. L'accès aux documents en cas de besoin sera simple grâce à leurs URIs indiqués dans les annotations.

L'exploitation des annotations sémantiques par des agents intelligents logiciels ou par des moteurs de recherche sémantiques disposant de capacité d'inférence vise à offrir une recherche « intelligente » sur le Web (externe ou interne) des documents répondant à la requête sémantique de l'utilisateur. Le système sera implémenté sous forme d'un système de plusieurs agents coopérants.

L'utilisation de l'ontologie dans le développement du système s'exprime dans toutes les composantes importantes :

- Pour l'interface utilisateur : Définition des besoins, expression des sujets de veille, formulation des requêtes sémantiques et visualisation des résultats ;
- Pour la création et la génération des annotations sémantiques ;
- Pour la recherche des documents externes ;
- Pour la définition des « wrappeurs » qui extraient les informations des sites Web semi-structurés.

Dans la partie qui suit, nous présentons un outil important dans le système, responsable pour la recherche sémantique, Corese.

5.2 CORESE

CORESE (Conceptual Resource Search Engine) [Corby, 2004] un moteur de recherche sémantique dédié au langage RDF (Resource Description Framework), langage standard du web sémantique proposé par le W3C. CORESE est basé sur les Graphes Conceptuels (GC) (Sowa, 1984). Ce moteur est dédié à des applications de web sémantique communautaire ou d'entreprise. CORESE permet de charger des ontologies représentées dans le langage RDF Schema, d'annoter les documents de cette mémoire en utilisant ces ontologies, et d'utiliser ces annotations pour rechercher des documents

5.2.1 Principes de CORESE

Grâce a l'adéquation possible entre RDF et les Graphes Conceptuels, CORESE combine les avantages de ces deux formalismes de représentation des

connaissances. RDF permet d'exprimer et d'échanger des méta-données dans un formalisme recommandé par le W3C. Les graphes conceptuels sont un modèle formel étudié depuis des années, ayant déjà fait ses preuves. Les GC fournissent un bon moyen d'expression, sont facilement lisibles, et des mécanismes de requêtes et d'inférences permettent de les manipuler.

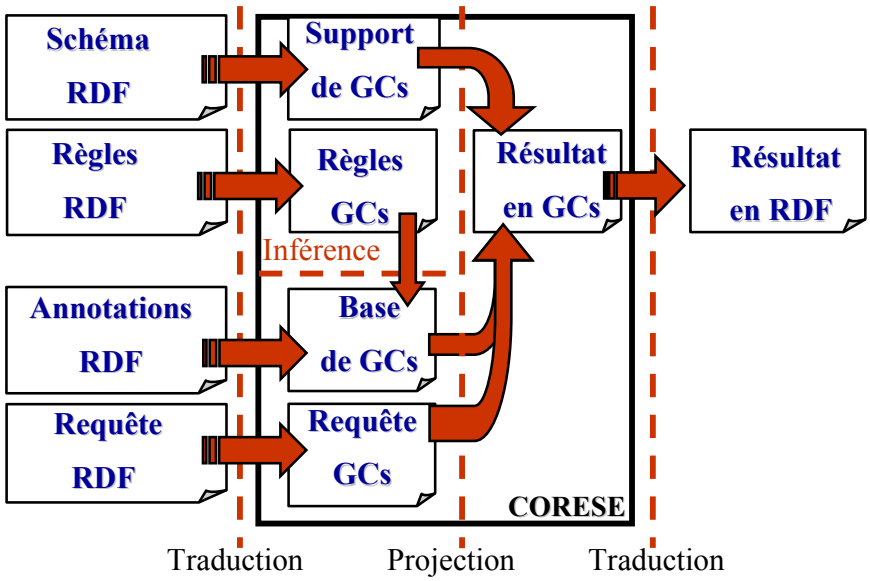


Figure 14 Principe de CORESE

CORESE est un moteur de recherche qui rend possible les inférences sur des annotations RDF, en les traduisant en GC puis restitue les résultats des GC vers du RDF. CORESE fait intervenir deux composantes : RDF(S) et GC. RDF permet de décrire le contenu des documents à travers les annotations sémantiques que l'on utilisera pour la recherche d'information. Les annotations sont basées sur une ontologie représentée en RDF Schema.

Comme on peut le voir sur la figure 14, la démarche de Corese repose sur la correspondance entre deux formats : RDF(S) et les GC. Les classes et les propriétés de RDFS sont traduites vers les types de concepts et les types de relations des GC (support) afin de pouvoir par la suite faire des requêtes dans la base RDF/GC. Une requête se présente comme un énoncé RDF, traduit en un graphe requête qui est ensuite projeté sur la base des GC afin de trouver les graphes qui s'apparient avec lui. Les graphes résultats sont ensuite traduits en retour vers RDF.

5.2.2 Traduction des modèles RDF(S) vers des GC

Le tableau 5.1 montre la correspondance entre RDF(S) et les graphes conceptuels (GC). Pour la déclaration d'une ontologie, nous avons le rapport suivant : les Classes RDF (rdf :Class) correspondent aux types de concepts en GC ; les propriétés RDF (rdf :Property, qui permet de déclarer les propriétés d'une ressource) correspondent aux types de relations de GC ; le *domaine* et le *codomaine* d'une propriété RDF ont leur équivalence en GC par la signature d'un graphe. Pour l'expression d'annotations, une *ressource* à décrire en RDF correspond à un Concept et les *propriétés RDF* correspondent aux relations GC. Enfin, les *annotations RDF* deviennent des graphes conceptuels.

RFS(S)	GC	
rdfs :Class rdf :Property domain, range	Type de concept Type de relation Signature	Pour la déclaration de l'ontologie
Ressource Ressource Anonyme Propriété Annotation RDF	Concept [Type de concept :référent] Concept générique Relation Graphe conceptuel	

Tableau 5.1 – Correspondance entre RDF(S) et les graphes conceptuels.

Afin d'enrichir une base d'annotations, CORESE propose des méthodes pour représenter et gérer des méta-propriétés (propriétés assignées aux relations RDF). Parmi ces méta-propriétés, on trouve la transitivité, la symétrie, la réflexivité et l'inverse. L'exemple suivant représente la propriété isMember qui est transitive et qui a comme relation inverse hasMember (une autre propriété)

```
<rdfs :Property rdf :ID= « isMember »>
       <cos:transitive>true</cos:transitive>
       <cos:inverse rdf:resource='#hasMember'/>
</rdfs:Property>
```

En utilisant le mécanisme de projection du formalisme des Graphes Conceptuels et sa propre manière de traiter les méta-propriétés, CORESE fait des raisonnements sur les GC pour améliorer la recherche d'informations et faire des inférences.

Pour interroger une base d'annotations, CORESE dispose d'un langage de requêtes prôche de SPARQL et basé sur des triplets en RDF et permet de manipuler des variables et des opérateurs : opérateurs arithmétiques, négation et comparaison de types.

CORESE dispose aussi d'un moteur d'inférences : en utilisant les concepts d'une ontologie et un ensemble de règles définies par défaut (transivité, symétrie, antisymétrie), il enrichit la base d'annotations permettant ainsi d'exploiter les connaissances implicites des annotations à travers des inférences. Ce moteur d'inférences applique ces règles à toutes les relations dans la base d'annotations ayant ces propriétés. Le module des inférences permet aussi de charger des règles différentes de celles définies par défaut, et qui établissent d'autres types d'inférences.

Avec CORESE, le système de veille peut offrir aux utilisateurs la capacité de faire des recherches sémantiques sur les annotations disponibles dans le système. CORESE est utilisé aussi pour les services concernant l'ontologie, fournir l'accès à des concepts, des propriétés nécessaires pour les algorithmes de recherche et annotation des documents.

5.3 Ontologies et agents sur le panorama du problème de veille au CSTB

Dans le système de veille technologique travaillant sur des informations hétérogènes, la tâche d'annotation sur les sources d'informations ou documents existants dans le système actuel ou déjà connus par le veilleur pourra être manuelle, semi-automatique ou automatique : cela dépendra de la nature des informations.

Mais le but n'est pas simplement d'obtenir des annotations sur les ressources internes de l'entreprise. Le système devrait être capable de découvrir, détecter sur le Web et rassembler des documents inconnus puis de les intégrer dans le système.

Donc, l'idée principale de notre solution est la suivante : quand un document sera recherché, le système essaiera tout d'abord de faire une recherche sémantique sur les ressources déjà annotées. Quant aux autres ressources du Web pertinentes pour

la requête mais non encore annotées, il est nécessaire de les découvrir sur le Web puis de les annoter et de stocker leurs annotations dans une base d'annotations, ces annotations devenant ainsi disponibles pour des requêtes ultérieures des utilisateurs.
La figure 15 montre le processus de veille avec la participation des agents et des ontologies.
Quand ils ont besoin de veille sur un certain sujet, les documentalistes disposent, après la phase de ciblage, d'une liste de thèmes qui correspondent à des concepts de l'ontologie.

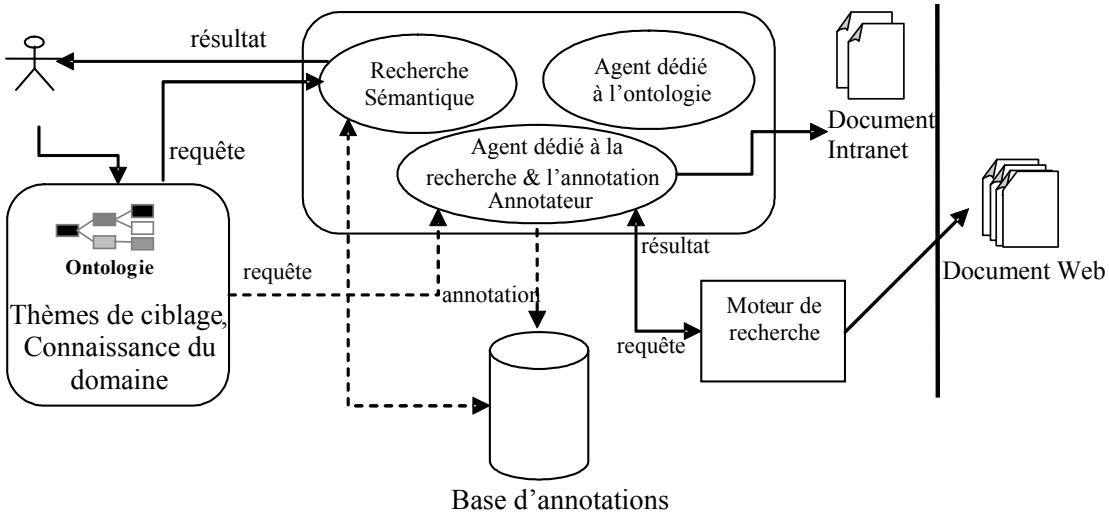


Figure 15 Ontologie et système multi-agents dans le système de veille.

Une requête est alors formulée et lancée sur ces thèmes. Deux cas peuvent se produire :

1. Les veilleurs connaissent déjà certaines sources qui peuvent contenir des informations utiles➔ Le système va donc les exploiter. Les agents vont traiter la requête et aller chercher les documents sur l'Intranet du CSTB ou sur des sites Internet.
2. La requête est envoyée à des moteurs de recherche comme Google. La masse des résultats renvoyés par ces moteurs est ensuite traitée par des agents spécifiques pour obtenir des résultats plus pertinents.

Dans ce système il faut distinguer la recherche des documents en deux étapes. Tout d'abord le système va effectuer une recherche sémantique utilisant CORESE sur les

bases d'annotations des documents disponibles. Puis, s'il n'obtient pas de résultat satisfaisant, une recherche classique sur les sources non encore annotées est effectuée, et dans ce cas, les documents constituant le résultat de la recherche devraient être analysés et traités pour que le système puisse rajouter les annotations sur ces documents dans la base d'annotations existante. Donc les deux tâches importantes à améliorer dans le processus de veille sont : la recherche de documents et l'annotation de documents. L'ontologie et les agents logiciels vont jouer un rôle important pour ces tâches.

5.4 Architecture du système

L'architecture du système s'articule autour de l'ontologie O'Watch. Tous les modules décrits ci-dessous sont implémentés sous forme d'agents logiciels. L'architecture du système OntoWatch intègre les modules suivants :

Un module de l'interface s'occupe des aspects utilisateurs, aide les utilisateurs à exprimer leur besoin de veille, visualiser et accéder aux annotations, documents, résultats de veille.

- Un gestionnaire de l'ontologie fournit les services ontologiques pour les autres modules en cas de besoin.
- Un composant s'occupe de la gestion des annotations dans le système, la recherche sémantique d'information reposant sur ces annotations. Ces deux composants utilisent certaines fonctions de CORESE.
- Un composant chargé de découvrir, chercher les documents externes pour alimenter la base d'annotations. Il comprend des modules plus fins correspondant à la tâche d'annotation aux niveaux différents :
  - le module « Wrapper » permet de générer semi-automatiquement des annotations RDF à partir des sources d'information structurée et semi-structurée.
  - le module « Recherche et Annotation » des documents externes utilise l'ontologie pour chercher sur le Web des documents répondant au besoin du veilleur et générer automatiquement des annotations RDF.

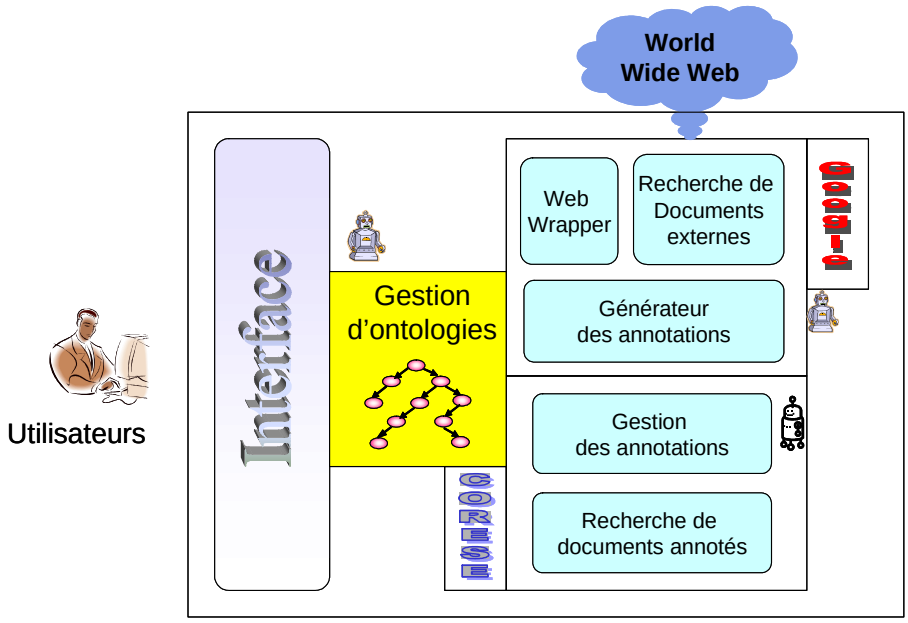


Figure 16 Architecture du système OntoWatch

5.5 Conclusion

Ce chapitre a donné une vision globale sur les composants principaux du système Onto Watch. La veille repose sur l'exploitation de la recherche sémantique grâce à l'aide du moteur Corese et de la recherche sur le Web, guidée par l'ontologie, pour générer des annotations sémantique. Les fonctionnements du système présentés à travers cette architecture seront concrétisés en paradigme multi-agents dans le chapitre 7.

6 Recherche et annotation des documents Web en utilisant l'ontologie

Jusqu'à présent la vision de Tim Berner-Lee sur le web sémantique n'est pas encore atteinte : on ne dispose pas encore des méta-données sur le contenu sémantique des ressources du Web et la recherche sur le Web est effectuée essentiellement par des requêtes basées sur les mots-clefs sur des portails tels que Google ou Yahoo. Ce chapitre décrit nos travaux exploitant des ontologies afin d'améliorer la recherche des documents sur le Web, et d'annoter ces documents. Tout d'abord, nous analysons l'apport de l'ontologie et puis nous présentons des algorithmes utilisant une ontologie pour rechercher et annoter les documents du Web en reposant sur différents critères : les branches d'un concept dans une ontologie et la distribution des descendants de concepts.

6.1 Apports de l'ontologie pour la recherche d'information sur le Web

Quand l'utilisateur fait la recherche sur le Web en utilisant des moteurs de recherche tels que Google et Yahoo, les résultats obtenus en réponse consistent en des dizaines de milliers de pages Web. L'utilisateur ne peut pas explorer toute l'information. Actuellement, beaucoup de moteurs de recherche utilisent une méthode d'indexation basée sur des mots contenus dans une page Web ou sur le nombre de liens vers les pages contenant les mots-clés. Par conséquent, les moteurs de recherche sont basés sur une approche qui classe les résultats selon la méthode simple de « ranking». Tandis que la vitesse est un avantage de ces moteurs, la pertinence est un autre problème. En comparaison avec les résultats empiriques selon lesquels les utilisateurs utilisent approximativement un à trois mots en moyenne [Croft et Cook, 1995], il est difficile de capturer l'intention initiale de recherche d'utilisateurs. Une requête pour "Java" a pu par exemple être une référence à l'île Java, au café Java ou au langage de programmation Java, selon l'intention de l'utilisateur. Pour surmonter ce problème, différentes techniques ont été conçues, par exemple l'expansion de requête [Qui et Frei, 1993] [Voorhees, 1994], où la liste de mots que l'utilisateur saisit est artificiellement augmentée avec son synonyme en utilisant des thesaurus tels que WordNet [WordNet, 2005].

Comme les moteurs de recherche aujourd'hui ne connaissent pas le contexte exact des entités à rechercher, une solution à ce problème est la recherche sémantique. Mais même dans le cas où l'utilisation d'un moteur de recherche classique basée sur les « mots-clés », est inévitable, une ontologie pourrait être utilisée pour obtenir de meilleurs résultats de recherche. Dans cette optique des ontologies ont été construites pour établir un vocabulaire commun dans certains domaines ainsi que des moteurs de recherche permettent l'accès focalisé sur des sujets spécifiques ont été proposé, il devrait donc être possible de combiner les points forts de ces deux approches.

Nous proposons d'employer des ontologies pour ajouter le contexte afin d'obtenir un ensemble de résultats plus significatif. Au lieu des mots-clés, des concepts de l'ontologie seront choisis pour former une requête. L'avantage est le fait que d'une part nous pouvons utiliser les concepts descendants des concepts initiaux dans l'ontologie pour constituer la requête ; l'ontologie aidera à réduire le manque de

pertinence dans les résultats de recherche obtenus à partir d'un moteur de recherche classique. Plus concrètement, grâce au vocabulaire défini dans une ontologie, l'algorithme de recherche peut recueillir des documents pertinents qui ne comprennent pas exactement le mot clef (correspond à un concept) original de l'utilisateur.

Par exemple, quand nous recherchons les documents concernant un "accident de voiture", il est nécessaire de pouvoir retrouver des documents évoquant non seulement "accident de voiture" mais également " accident de camion", "collision de voiture", "explosion de bus". Pour cette raison la requête doit pouvoir inclure toutes ces alternatives plus précises que "accident de voiture".

D'autre part, l'expansion de requête par les concepts descendants dans l'ontologie permet de mieux spécifier le contexte de recherche. Par exemple, la réponse du moteur de recherche à une requête avec le concept (correspond à un mot clef) « Grue » peut comprendre un document concernant un animal, alors que le veilleur veut collecter des documents relatifs à une machine. Mais avec une requête avec ses concepts descendants tels que « Grue Hydraulique », la possibilité d'obtenir des documents non pertinents est réduite.

Le problème est comment former la requête à envoyer au moteur de recherche.

6.2 Stratégie d'annotation des documents du Web

Dans le scénario de la veille au CSTB, nous visons à découvrir de nouvelles informations ou connaissances dans le domaine du bâtiment et de la construction. Ainsi, notre système doit pouvoir produire des annotations permettant de savoir qu'un document donné est associé à un sujet donné et est utile pour un tel utilisateur (resp. groupe).

Voici quelques cas spécifiques où l'annotation automatique d'un document textuel est possible :

- Annotations correspondant aux méta-données Dublin-Core comme : le titre, les auteurs, le type de document, la date de création, etc.
- Annotations associées à des concepts non évoqués directement dans le texte : par exemple, pour un document concernant le sujet de la "gestion des

connaissances", l'observateur peut présenter une annotation indiquant que ce document est intéressant pour les ingénieurs du service SIA du CSTB. Le moteur de recherche sémantique CORESE utilise des règles d'inférence pour produire automatiquement ce type d'annotations (Corby et al. 2002,2004,2006).

- Pour les annotations dont le contenu est extrait à partir des textes, il est nécessaire de partir du cas le plus simple au cas le plus compliqué pour reconnaître :
    - les mots significatifs : mots gras, italiques, entre guillemets…
    - les mots particuliers qui associent au type de document sa nature...
    - les termes correspondant aux concepts de l'ontologie.

La manière de les inclure dans l'annotation dépendra du point de vue de l'annotateur et sera prédéfinie pour la génération automatique d'annotation.

Si l'information à extraire pour générer l'annotation est de type structuré ou semi-structuré, l'utilisation de règles XSLT (en utilisant la puissance d'expression du langage de chemin d'accès XPATH) pour la génération semi-automatique d'annotations (Cao et al. 2003) est pertinente et peu coûteuse.

6.3 Algorithme général

Dans cette section, nous présentons notre approche générale développée dans un algorithme permettant d'envoyer au moteur de recherche Google une requête formée par des concepts dans l'ontologie et puis d'annoter automatiquement les documents trouvés dans les résultats obtenus. Google est actuellement considéré comme le moteur de recherche le plus efficace avec plus de deux milliards de pages Web indexées. Les développeurs peuvent utiliser le service de recherche de Google dans leur application, grâce à Google Web API. Le cas idéal de l'algorithme est celui où une requête de l'utilisateur, considérée comme une liste de concepts dans l'ontologie, pourrait être remplacée par une requête générée par le système, et comprenant tous les concepts descendants de tous les concepts initiaux de la requête de l'utilisateur. A partir des labels correspondant à chaque concept de l'ontologie

dans la requête système, l'algorithme peut formuler une requête (en chaîne de caractères) pour faire la recherche sur le Web avec Google.

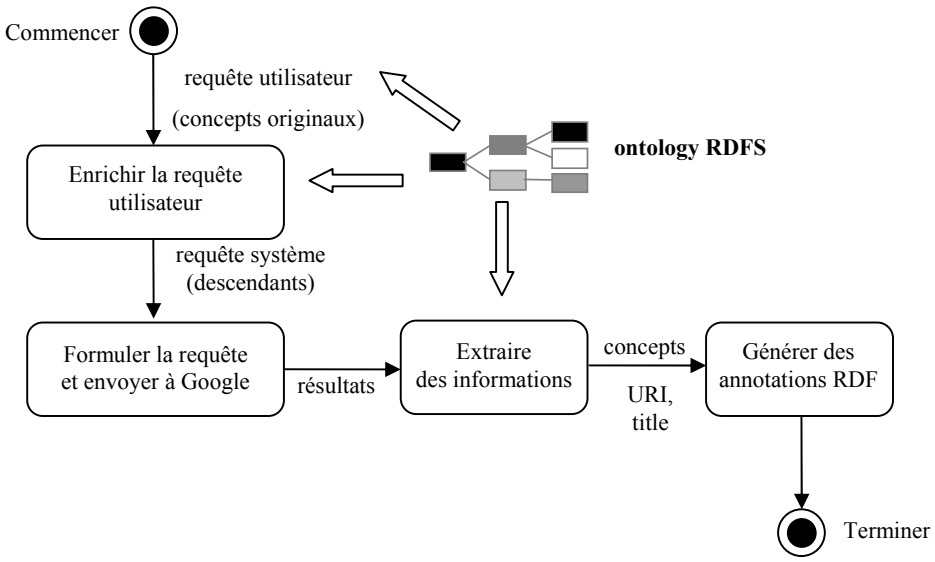


Figure 17 Principe de l'algorithme général

Pour chaque requête, le résultat obtenu à partir de Google est une liste d'éléments spécifiques représentant chaque document trouvé. A partir de ces éléments, nous pouvons obtenir : l'URL du document, le titre, un texte extrait à partir du document montrant les mots-clés de la requête dans le contexte où ils apparaissent dans le document, et d'autres informations. Ainsi nous pouvons traiter ce texte pour extraire des mots-clés trouvés dans le contenu du document.

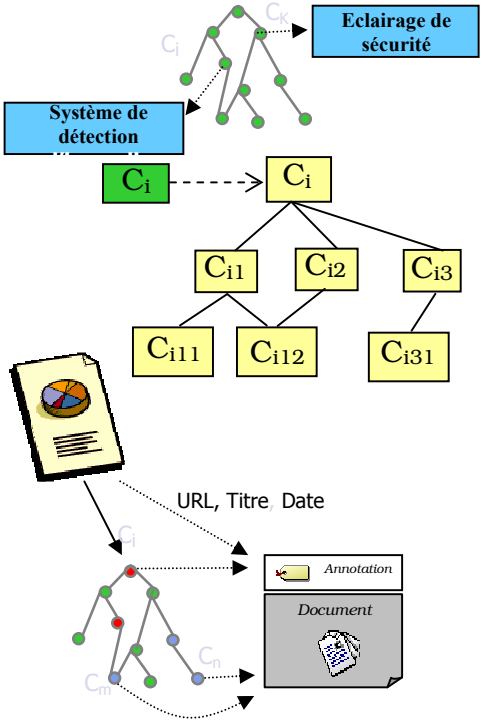
6.3.1 Description de l'algorithme

NotonsVu $Request_U$ la requête de l'utilisateur envoyée à Google sous la forme d'un ensemble de n concepts dans l'ontologie et $Request_S$ est la requête que le système générera à partir des concepts originaux de $Request_U$. Dans ce qui suit, pour simplifier, nous appellerons « concepts utilisateur » les concepts initiaux figurant dans la requête de l'utilisateur et « concepts système » ceux figurant dans la requête générée par le système pour être envoyée à Google.

L'utilisateur va tout d'abord exprimer une requête initiale se composant de mots-clés correspondant à une liste de concepts utilisateur $\{C_i\}$ de l'ontologie. Pour

simplifier, nous supposerons que ces concepts sont indépendants, c'est-à-dire il n'existe pas :

C_j et C_k tels que $C_i \in Request_U$ et $C_k \in Request_U$ et C_j est un sous-concept de C_k.



a) $Request_U = \{C_i\}$ i=0,1,… n

Pour pouvoir obtenir le plus possible de ressources dont le contenu pourrait concerner les thèmes spécifiés dans la requête de l'utilisateur, il faut ajouter systématiquement dans la requête initiale de l'utilisateur les mots-clés correspondant aux sous-concepts des concepts utilisateur C_i dans l'ontologie.

On définit Descendants(C_i) comme un ensemble contenant tous les descendants de C_i dans l'ontologie.

La requête idéale que le système va envoyer à Google à partir de $Request_U$ comprendra donc tous les concepts descendants de chaque $C_i \in Request_U$ dans l'ontologie O.

$Request_S = \{C_j$ tel que $\exists\, C_i \in Request_U$ et $C_j \in \{Descendants\ (C_i)\}\}$

Pour obtenir $Request_S$, pour chaque $C_i \in Request_U$, i= 0,1,… n on prend D_i = Descendants(C_i)

b) $Request_S = \cap_{i=0..n} \{D_i\}$, i= 0..n. = Descendants $(C_i) \cap$ Descendants $(C_j) \cap$ …. Descendants (C_n)

c) Constituer la requête (une chaîne de caractères) avec tous les labels des concepts de $Request_S$. Les opérateurs insérés entre ces labels sont « ET » et « OU ». Normalement, le veilleur souhaite une réponse concernant tous ses concepts initiaux correspondant à tous les thèmes ciblés dans sa veille. L'opérateur « OU » est utilisé entre les labels de concepts descendants d'un concept utilisateur. Par contre

l'opérateur « ET » est inséré entre les groupes de concepts correspondant aux concepts utilisateurs.

A partir de l'étapes d) jusqu'à la fin, l'algorithme effectue une tâche particulière : faire la recherche par Google avec une requête String et puis générer des annotations. Cette tâche est répétitive dans les algorithmes donc pour la réutilisation nous l'encapsulons dans un module BasicSearchAnnotate(S, $Request_U$) pour la réutilisation ultérieure.

d) Lancer directement la recherche avec cette requête par Google.

e) Prendre les résultats de Google et analyser les structures attachées aux documents retournés dans le résultat de Google pour annoter le document. En effet, chaque résultat de Google fournit : l'URL du document trouvé, le titre, la date et enfin l'ensemble des mots-clés de la requête trouvés dans le document. Nous avons donc deux options :

- Soit annoter le document avec tous les concepts correspondants aux mots clés indiqués dans le résultat de Google concernant ce document..
- Soit extraire les concepts les plus précis (c'est-à-dire le plus bas dans la hiérarchie de concepts de l'ontologie), et annoter le document avec ces concepts ainsi qu'avec les C_i dans $Request_U$ dont ils sont descendants. Pour cela, nous allons procéder comme suit. Nous avons en entrée d'une part, $\{C_i\}$ de $Request_U$, d'autre part, pour chaque ressource résultat, $\{C_k\}$ correspondant aux mots-clés trouvés dans cette ressource résultat. La sortie S devrait comprendre une liste de concepts pour annoter le document : cette sortie est construite comme suit :
  - Pour chaque $C_l \in \{C_k\}$:
    - Si $C_l \in Request_U$, ajouter C_l dans S, $\{ C_k \} = \{ C_k \} - C_l$.
    - Sinon prendre D_l = Descendants(C_l) en supposant que ces descendants sont rangés par ordre de profondeur décroissante.
      - S'il existe $C_{l'} \in \{C_k\} \cap D_l$, ajouter $C_{l'}$ dans S, et $\{ C_k \} = \{ C_k \} - C_{l'}$
      - Sinon ajouter C_l dans S, $\{ C_k \} = \{ C_k \} - C_l$

- Annoter le document avec les concepts appartenant à S et les autres informations extraites.

Remarque:

Cet algorithme repose sur l'hypothèse simplificatrice qu'étant donné un mot-clé, nous ne pouvons trouver qu'un seul concept correspondant à ce mot-clé dans l'ontologie. Autrement dit, il n'existe pas dans l'ontologie deux classes ayant le même label.

L'algorithme fonctionne sur l'hypothèse qu'il n'y a pas de limite sur le nombre de mots-clefs dans une requête système: $Request_S$. Cependant, le moteur classique sur lequel nous avons choisi de reposer (du fait du grand nombre de pages indexées et du fait de son API utilisable par les programmeurs) est Google ; or Google limite le nombre de mots-clés d'une requête alors que, malheureusement, dans la plupart des situations réelles, dans une ontologie assez grande, le nombre de concepts descendants dépasse cette limite. C'est la raison pour laquelle l'algorithme général est applicable seulement dans le cas où $Card(Request_S)$ est inférieur à cette limite. Dans le cas contraire, pour surmonter ce problème, nous avons développé des algorithmes dont l'idée principale est : au lieu de générer une requête système dont le nombre de concepts (correspondant aux mots-clés) dépasse la limite de Google, nous générons plusieurs petites requêtes système dont le nombre de concepts respecte la limite de Google. En fonction de la manière de choisir les concepts descendants pour formuler ces petites requêtes, ces algorithmes sont divisés en deux types :

- Les algorithmes basés sur les branches de l'ontologie,
- Les algorithmes basés sur la distribution équilibrée entre des descendants de concept utilisateur.

6.4 Algorithmes basés sur les branches de concept utilisateur

Définition : *Une branche de l'ontologie ayant C_i comme concept racine est l'ensemble de tous les concepts successifs apparaissant dans un chemin à partir du concept racine C_i jusqu'à n'importe quel concept feuille descendant de C_i.*

Par exemple : une branche de l'ontologie correspondant à C_i dans la figure 18 peut être : (C_i, C_{i1}, C_{i11}) ou (C_i, C_{i3}, C_{i31}), …

Le principe des algorithmes basés sur les branches de l'ontologie est : au lieu de interroger le Web avec tous les descendants de concepts utilisateurs dans une seule requête, envoyer plusieurs requêtes. Chacune correspond aux concepts dans une branche de chaque concept utilisateur. La différence dans la manière d'utiliser les branches de concepts de l'ontologie pour faire la recherche d'information et traiter les résultats de recherche pour générer des annotations mène à deux algorithmes distincts.

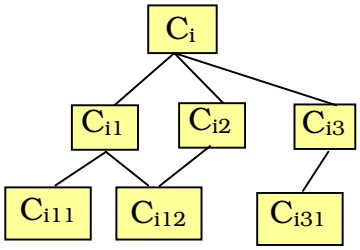


Figure 18 Concept C_i et ses descendants

6.4.1 Premier algorithme : interroger le Web avec toutes les branches de concepts utilisateurs dans la requête initiale.

Le premier algorithme utilise toutes les branches de l'ontologie à partir de chaque concept dans la requête de l'utilisateur afin de former plusieurs requêtes à envoyer au moteur de recherche Google. Chaque branche correspond à une requête. Le traitement des documents dans le résultat de chaque requête reste presque le même que celui effectué dans l'algorithme général.

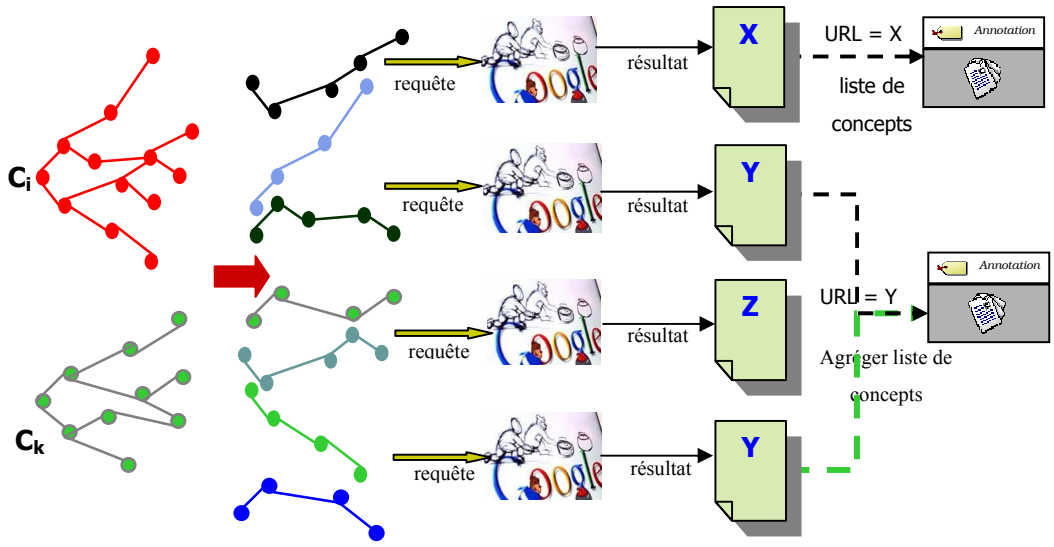


Figure 19 Recherche avec toutes les branches des concepts utilisateurs

Après avoir reçu toutes les annotations pour chaque branche, le travail restant est de :

- Eliminer les redondances, pour éviter les annotations identiques générées par l'algorithme à cause des intersections entre des branches. C'est le cas où un document est annoté plusieurs fois avec un même ensemble de concepts.
- Agréger tous les concepts extraits à partir des différentes branches concernant un même document.

Description de l'algorithme

a) $Request_U = \{C_i\}$ i=0,1,… n

b) Pour chaque C_i :

- calculer toutes les branches de C_i : Branches(C_i)
- ajouter Branches(C_i) à l'ensemble global BRAN.

c) Pour chaque branche Br_i dans BRAN :

- Constituer la requête S (une chaîne de caractères) avec tous les labels de concepts du Br_i. Utiliser l'opérateurs OU entre les labels de concepts dans Br_i.
- Faire la recherche par Google et générer des annotations : appeler BasicSearchAnnotate (S, $Request_U$)

d) Eliminer les redondances et agréger des listes de concepts :

- Chercher les annotations décrivant un même document (même URI)

- Si elles sont identiques, éliminer les redondances
- Si non, agréger toutes les listes de concepts dans ces annotations dans une liste globale de concepts et générer une seule annotation correspondant à ce document.

6.4.2 Deuxième algorithme : Recherche avec une branche

Dans le deuxième algorithme, nous effectuons seulement la recherche initiale avec une branche choisie aléatoirement. Pour chaque document figurant dans les résultats obtenus à partir de Google, nous recherchons dans les autres branches de l'ontologie, les concepts qui sont appropriés au contenu de document. Mais, au lieu de parcourir tous les concepts restants de l'ontologie afin de les comparer au contenu de ce document, dans cet algorithme, nous lançons une recherche supplémentaire focalisée sur le site Web contenant ce document pour voir si ce même document est également trouvé avec une autre requête correspondant à une autre branche de l'ontologie. Si c'est le cas, nous pouvons annoter ce document avec tous les concepts de ces branches.

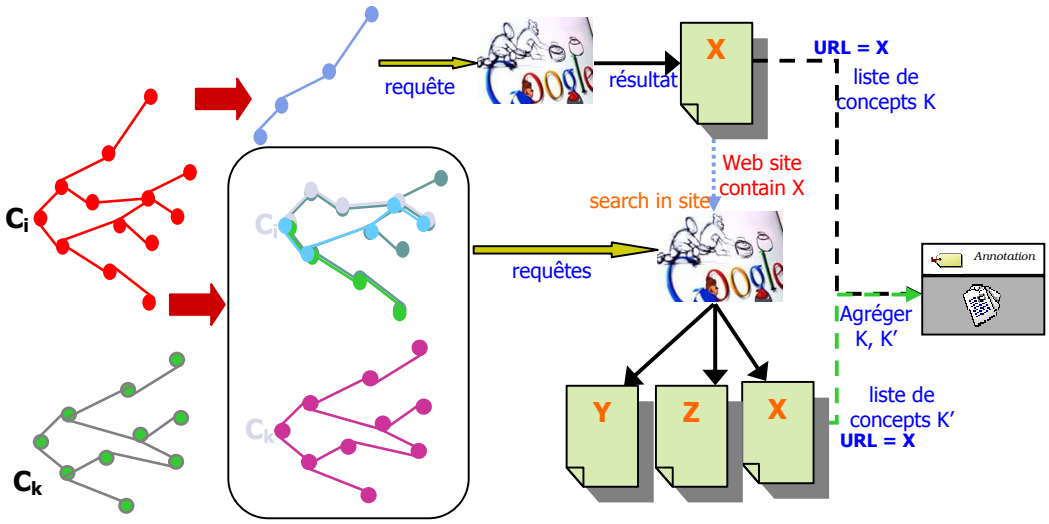


Figure 20 Recherche supplémentaire dans le site pour agréger les concepts dans les différentes branches

Description de l'algorithme

a) $Request_U = \{C_i\}$ i=0,1,… n

b) Pour chaque C_i :

- calculer toutes les branches de C_i : Branches(C_i)
- ajouter Branches(C_i) à BRAN.

c) Choisir une branche Br dans BRAN.

d) Constituer la requête S (une chaîne de caractères) avec tous les labels de concepts du Br. Utiliser les opérateurs OU entre les labels de concepts dans Br.

e) Lancer directement la recherche avec cette requête sur Google.

f) Prendre les résultats de Google, et analyser les structures attachées aux documents retournés dans le résultat de Google, pour annoter le document. En effet, chaque résultat fourni par Google : l'URL du document trouvé, le titre, la date et enfin l'ensemble des mots-clés de la requête trouvés dans le document.

g) Pour chaque élement D dans le résultat de Google faire :

- Obtenir URLSite = URL du site web contenant D.
- Pour chaque Br'≠ Br dans l'ensemble de branches restant dans BRAN : Chercher par Google avec Requête = l'ensemble des concepts de Br' avec le paramètre website= URLSite
- Pour chaque élément D' dans le résultat de Google faire
  si D.URL = D'.URL alors
  Obtenir K'= liste des concepts trouvés dans le document D'.
  Ajouter à l'ensemble Ann le concept le plus profond de Br appartenant à K'

finpour

finpour

Annoter D avec tous les concepts dans Ann. Retourner D avec ses annotations.

6.4.3 Exemple illustrant les deux algorithmes

Supposons que l'utilisateur demande des documents concernant deux concepts de l'ontologie O'Watch : C_h: "système de détection d'incendie" et C_k:"éclairage de sécurité". Tous leurs concepts descendants dans l'ontologie sont montrés dans la figure 21.

Le nombre de concepts descendants de C_h et de C_k est 14 et dépasse la limite de Google (10).

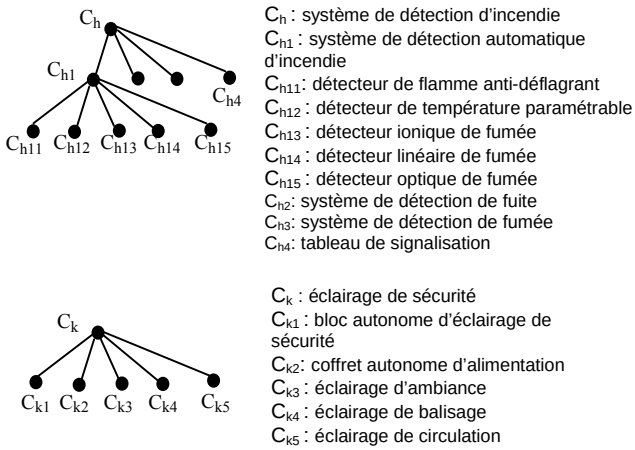


Figure 21 Concepts initiaux avec leurs concepts descendants.

Dans la première solution, notre algorithme enverra à Google des requêtes de toutes les branches de l'ontologie à partir de C_h ou de C_k : (C_h, C_{h1}, C_{h11}), (C_h, C_{h1}, C_{h12}), …(C_h, C_{h1}, C_{h15}),….., (C_h, C_{h2}),… (C_h, C_{h4}), (C_k, C_{k1})... (C_k, C_{k5}).

Dans la deuxième solution, du fait que chaque document D a été trouvé avec une requête correspondant à une branche choisie aléatoirement ($C_{h,}$,C_{h1},C_{h13}) par exemple, l'algorithme recherchera dans le site Web contenant D, avec des requêtes correspondant à d'autres branches : (C_h, C_{h1}, C_{h11}), (C_h, C_{h1}, C_{h12}), (C_h, C_{h1}, C_{h14}),….., (C_k, C_{k5}),... Si avec une branche telle que (C_k, C_{k4}), le document D est trouvé encore une fois dans le résultat de la recherche, D sera annoté avec ($C_{h,}$,C_{h13}, C_k,C_{k4}).

6.5 Algorithmes basés sur la distribution équilibrée entre des descendants de concepts

Notre évaluation des deux algorithmes précédents a montré qu'ils sont plus appropriés pour obtenir des documents spécialisés que des documents généraux. Cependant ils souffrent de quelques inconvénients :

- Premier algorithme : les documents trouvés sur le Web dans les résultats en réponse à toutes ces requêtes sont alors automatiquement analysés pour

éliminer les redondances éventuelles. Dans le cas où le même document est trouvé avec des requêtes correspondant à des branches différentes, les concepts de ces branches sont agrégés pour obtenir une liste complète de concepts pour annoter le document. Mais si les domaines liés aux différentes branches sont relativement indépendants, il sera peu probable de trouver un document concernant plusieurs de ces branches : le système risque alors de manquer les documents appropriés liés à tous les concepts utilisateur. Finalement, il faudra traiter davantage de redondances quand deux branches ne diffèrent que d'un seul concept.

- Le deuxième algorithme favorise trop une seule branche de l'ontologie car il dépend du choix aléatoire d'une branche pour la première requête. Les documents trouvés sont d'abord annotés avec des concepts de cette branche, puis le système essaye d'ajouter d'autres concepts issus des autres branches. Contrairement au premier algorithme, si le vocabulaire dans cette branche est trop spécialisé et distinct des autres branches, le système risque de manquer des documents importants concernant les sujets indiqués par ces autres branches.

La difficulté majeure rencontrée dans l'algorithme utilisant une ontologie pour rechercher sur le Web et produire des annotations est que le nombre de concepts descendants est souvent trop grand, et il n'existe pas une solution parfaite si nous ne voulons pas perdre aucun concept.

Revenons au besoin de l'utilisateur : Quand il veut chercher les documents concernant le thème abordé avec le concept C, peut-être veut-il d'abord obtenir les documents correspondant aux concepts qui sont assez solidaires de C (i.e dont la distance sémantique n'est pas grande). Par exemple si nous cherchons des documents concernant « Knowledge Management », il espère obtenir les documents concernant « Knowledge Modelling » et « Knowledge accquisition » plutôt que les documents concernant « OWL »

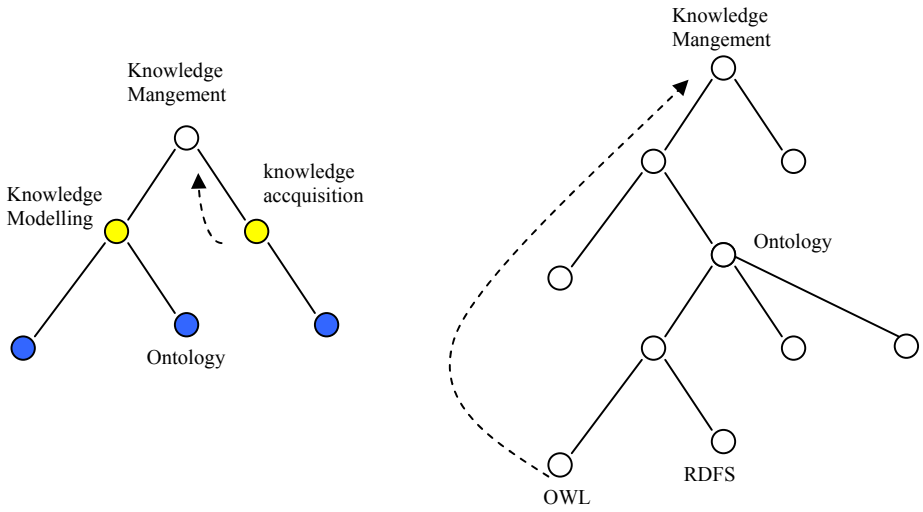


Figure 22 Les rapports entre les concepts au différent niveau de profondeur

Dans ce cas où nous ne pouvons pas utiliser en même temps tous les concepts descendants, il faut choisir les concepts descendants à un niveau pas trop profond. Dans le cas où l'on s'intéresse aux concepts très spécialisés, peut-être le veilleur connaît-il le concept plus général mais assez près de ces concepts dans la hiérarchie de l'ontologie. Par exemple, si le veilleur pense à OWL ou RDFS, il mettra « ontologie » dans la requête originale.

Pour cette raison, nous avons cherché une autre solution qui assure mieux que le document trouvé par Google est lié à tous les concepts utilisateur. Plus précisément, les requêtes produites par le système, qui se composent d'une liste de concepts descendants, ne doivent favoriser aucun concept utilisateur au détriment des autres.

6.5.1 Principe de l'algorithme

Nous avons développé un algorithme permettant de rechercher et d'annoter des documents qui concernent le plus possible de concepts utilisateur. Ainsi la solution de compromis que nous avons choisie ne vise pas à prendre le plus possible des concepts descendants de chaque concept utilisateur, mais tient compte plutôt d'une distribution équilibrée entre les différentes branches issues des concepts utilisateur. Comme le nombre de concepts autorisés dans une requête de Google est trop petit par rapport au nombre des concepts descendants pouvant être sélectionnés, notre

algorithme doit faire des choix multiples. Autrement dit, plusieurs requêtes correspondant aux sélections diverses des concepts sont générées.

Pour assurer la distribution équilibrée entre les différentes branches, nous reposons sur un critère : le nombre de concepts descendants choisis sur une branche va dépendre du poids de leur concept initial par rapport aux autres concepts utilisateur. Soit Total_Desc le nombre de tous les concepts descendants d'au moins un concept utilisateur, et Local_Desc (C) le nombre de concepts descendants du concept initial C. Le poids de C est la valeur de Local_Desc (C)/Total_Desc. La présence dans la requête générée d'au moins un concept descendant pour chaque concept utilisateur évite les inconvénients des deux algorithmes précédents. Pour chaque concept utilisateur, nous avons un nombre limite de concepts descendants à choisir afin de contribuer à la requête finale.

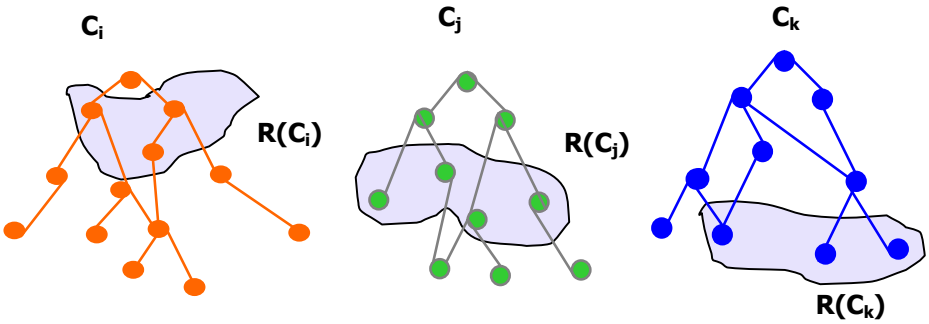


Figure 23 Distribution des descendants des concepts utilisateur dans une requête système

Cependant, notons un cas d'exception : si le nombre de concepts descendants d'un concept utilisateur est trop petit par rapport aux autres concepts utilisateurs, le nombre limite pour ses concepts descendants est égal à 1, et donc la requête système peut ne pas prendre en compte ce concept. Or dans l'ontologie, si un concept utilisateur n'a qu'un ou deux sous-concepts descendants, ces sous-concepts jouent souvent un rôle (presque) aussi important que leur parent et ils sont très proches sémantiquement du concept utilisateur. Par conséquent il serait regrettable de les manquer dans des requêtes systèmes. Dans ce cas-là, l'algorithme prend tous ces concepts descendants dans la requête système. Si c'est le cas commun pour tous les

concepts utilisateur, l'algorithme utilise non seulement les concepts descendants mais aussi leurs synonymes pour générer la requête système.

Le problème est la stratégie de sélection des concepts descendants. Contrairement à nos algorithmes précédents, qui choisissent des concepts en profondeur, pour un concept utilisateur, le nouvel algorithme choisit ses concepts descendants en largeur. Le résultat de ce processus est un ensemble de requêtes partielles, correspondant à chaque concept utilisateur. La combinaison de ces requêtes partielles nous donne l'ensemble final des requêtes système.

6.5.2 Description de l'algorithme

Le module principal *BalancedOntologySearchAnnotation* est responsable de la recherche des documents sur le Web par le moteur de recherche Google avec un ensemble de requêtes générées par l'algorithme. Ensuite, à partir de chaque document figurant dans les résultats, les termes correspondant aux concepts de l'ontologie sont extraits et enregistrés dans une table de hachage dont la clé est l'URL du document, afin de générer une annotation RDF sur ce document. En reposant sur la comparaison des URLs de chaque document, on élimine toutes les redondances et le système va agréger les listes de concepts quand un même document a été trouvé par les requêtes différentes. Le module prend en entrée la requête de l'utilisateur $Request_u$ (qui est en fait une liste de concepts) et génère des annotations représentées en RDF.

```
Algorithme BalancedOntologySearchAnnotation(Requestu)
Q = GenerateQueries (Requestu)
// Ann : table de hachage dont la clé est l'URL de chaque document trouvé.
Ann ← ∅
pour chaque requête q ∈ Q faire
      Envoyer q à Google
       pour chaque document D dans les résultats de Google faire
         K ← ExtractConcept(D)
      si D.URL n'était pas dans Ann
        Ajouter K et URL de D à Ann
      sinon Mettre à jour l'ensemble des concepts à annoter avec D.URL
      finsi
      finpour
finpour
Annoter tous les documents D dans Ann avec leurs URL et l'ensemble de
concepts attachés.
```

La partie la plus importante est l'algorithme générant en utilisant l'ontologie, des requêtes système à partir d'une requête de l'utilisateur. Tout d'abord, comme nous l'avons mentionné, pour assurer qu'une requête système générée ne favorise aucune branche particulière de l'ontologie, le système calcule, pour chaque concept utilisateur, le nombre limite de concepts descendants pouvant apparaître dans la requête système. Cette limite repose sur le rapport entre le nombre de concepts descendants de ce concept utilisateur et le nombre total de concepts descendants des concepts utilisateur.

Algorithme ConceptNumberLimit($Request_u$)
Desc = Ensemble de tous les concepts descendants des concepts figurant dans $Request_u$
S ← Card(Desc)
pour chaque $C_i \in Request_U$ **faire**

$$Limit_i \leftarrow \frac{Card(Descendant(C_i))}{S} * Google\_limit$$

finpour

En effet, dans cet algorithme, la requête système se compose de certaines requêtes partielles. Chaque requête partielle est représentée par une liste de concepts descendants d'un concept utilisateur. Pour chaque concept initial, il y a plusieurs requêtes partielles correspondantes pouvant être choisies par un algorithme décrit dans le module SelectPartialConcepts. Ainsi le résultat du module GenerateQueries est un ensemble de toutes les combinaisons possibles de toutes les requêtes partielles de chaque concept utilisateur.

Algorithme GenerateQueries($Request_u$)
pour chaque concept $C_i \in Request_U$ **faire**
Obtenir l'ensemble de toutes les requêtes partielles générées à partir de Ci. Chaque requête partielle est un ensemble de concepts choisis parmi les concepts descendants de C_i:
Q_i ← SelectPartialConcepts(C_i, $Limit_i$)
finpour
Q = Combiner chaque requête partielle dans toutes les Q_i afin de générer l'ensemble global de requêtes utilisant l'algorithme de combinaison.

Nous savons que le nombre de concepts descendants de chaque concept dépasse souvent la limite autorisée. Ainsi le problème est d'avoir une bonne stratégie pour

sélectionner, parmi ces concepts, les meilleurs concepts pour former les requêtes partielles. L'idée principale de l'algorithme *SelectPartialConcepts* est la suivante :

Calculer le niveau maximum de profondeur issu du concept utilisateur dans l'ontologie. Considérant une ontologie comme un graphe, nous définissons le niveau maximum de profondeur du concept C comme le nombre de concepts (C exclu) dans le chemin de C à son concept descendant le plus profond C_d. Tous les concepts fils directs de C sont au niveau de profondeur 1, et ainsi de suite. Soit K_i le niveau maximum de profondeur du concept utilisateur Ci ; le nombre de requêtes générées est : $\prod_{i=1..n} K_i$, où n est le nombre de concepts de l'utilisateur.

À chaque niveau de profondeur issu de C, le système choisit un certain nombre de concepts, pour constituer une requête partielle (en respectant la limite imposée).

Les entrées de ce module sont le concept utilisateur C et le nombre de concepts descendants qu'il est autorisé de choisir. La sortie est l'ensemble de requêtes partielles générées.

```
Algorithme SelectPartialConcepts(C, Limit)
G ← ∅
k ← MaxDepthLevel(C)
pour i = 1 to k faire
 q ← SelectConceptFromLevel(C, i, Limit)
 Ajouter q à G
finpour
```

Le module *SelectConceptFromLevel* est responsable du choix d'un nombre prédéfini de concepts à partir d'un certain niveau de profondeur. Le système commence par prendre tous les concepts *SetC* au niveau actuel *level* de profondeur. Si le nombre de concepts à ce niveau est plus petit que la limite indiquée, le système passe au prochain niveau de profondeur et ainsi de suite.

```
Algorithme SelectConceptFromLevel(SetC, level, Limit)
Entrée
Begin
      Laisse k ← Card(SetC)
      si k ≥ Limit alors
       Ajouter un nombre de Limit premiers concepts dans SetC à OutC.
       Exit
      sinon si k < Limit alors
```

```
        Ajouter tous concepts de SetC à OutC.
        NextSet← GetAllConceptsNextLevel(SetC)
        SelectConceptFromLevel(NextSet,level+1,Limit – k)
       finsi
End
```

Revenons au exemple dans la figure 21. Supposons que l'utilisateur demande des documents concernant deux concepts de l'ontologie: C_h: "système de détection d'incendie" et C_k:"éclairage de sécurité". Grâce à l'aide de l'interface du système, l'utilisateur peut naviguer dans l'ontologie et trouver facilement des concepts représentant les sujets de veille. L'algorithme va calculer le nombre limite des concepts descendants de C_h et de C_k à choisir, qui sont l_h et l_k. Puis pour le concept C_h comme son niveau de profondeur est 2 donc : l'algorithme va choisir des groupes de l_h concepts descendants à partir du niveau 1 par exemple (C_{h1}, C_{h2}, … C_{h1i}) et à partir de niveau 2 par exemple (C_{h11}, C_{h12}, … C_{h1i}). Pour C_k le niveau de profondeur est 1, donc la sélection de l_k concepts est effectuée seulement parmi ses fils : C_{k1}, C_{k2},.. C_{kj}. La combinaison de tous ces ensembles de concepts donnera finalement un ensemble des requêtes globales à envoyer à Google

6.6 Extension de l'algorithme avec la prise en compte des synonymes

Nous avons trouvé une situation où tous les trois algorithmes ne sont pas très efficaces pour l'amélioration de la recherche d'information du veilleur, comme nous ne pouvons pas exploiter la relation généralisation-spécialisation entre les concepts. Les concepts utilisateurs spécifiés n'ont pas ou très peu de descendants. Pour résoudre cette difficulté, une extension de l'algorithme a été réalisée, pour prendre en compte tous les synonymes de mots-clefs définis dans le concept utilisateur.
Le principe de l'algorithme est le suivant:

a) $Request_U = \{C_i\}$ i=0,1,… n
b) Remplacer $Request_U$ pas la requête système $Request_S$ composée de tous ces concepts descendants.
c) Enrichir $Request_S$ avec tous les synonymes de ses concepts, on obtiendra la $Request_{Syn}$

d) Formuler une requête S en chaîne de caractères à partir des labels de concepts inclurant les synonymes dans la $Request_{Syn}$. Les opérateurs « ET » et « OU » sont insérés entre ces mots-clefs d'une façon appropriée.
e) Envoyer cette requête S à Google et générer des annotations comme les algorithmes décrits plus hauts.

6.7 Conclusion

Utiliser des vocabulaires dans une ontologie pour enrichir la requête d'utilisateur est une approche nouvelle. Avec ces algorithmes, nous profitons tout d'abord de la puissance de Google pour la recherche sur le Web pour améliorer la pertinence des résultats obtenus grâce à l'utilisation de l'ontologie. Bien que nos algorithmes exploitent la relation de subsomption entre les concepts dans l'ontologie, les ontologies manipulées ne sont pas nécessairement des arbres. Grâce au moteur Corese qui offre une traduction de RDF(S) vers les graphes conceptuels, les algorithmes traitent l'ontologie comme un support dans le formalisme des graphes conceptuels, et tiennent compte du problème de multi-héritage entre les concepts.

Les algorithmes que nous avons présentés dans ce chapitre n'exploitent que la relation de subsomption entre les concepts dans une ontologie et ses synonymes, mais ils améliorent de manière remarquable le résultat de recherche des documents sur le Web externe.

Nous pouvons encore chercher à exploiter les autres aspects ontologiques pour faire évoluer ces algorithmes, par exemple les relations autre que la subsomption entre les concepts, les instances.

7 Architecture multi-agents pour le système de veille

Ce chapitre présente la conception rationnelle que nous avons suivi pour obtenir une architecture multi-agents aidant au scénario de la veille technologique et scientifique envisagé. Nous nous sommes intéressés en particulier à la recherche d'information sur le Web et à la génération des annotations sémantiques.

Nous allons tout d'abord expliquer chaque étape principale de l'analyse descendante des fonctionnalités du système. Nous verrons que par l'identification des sociétés dédiées à certaines ressources de la tâche de veille et l'analyse de leur organisation, nous descendons au point où des rôles et leurs interactions peuvent être identifiées. Puis, nous allons décrire les caractéristiques et la documentation des rôles et des protocoles soutenant chaque sous-société dédiée.

7.1 Conception d'une société d'agents pour le système de veille

Lorsqu'il envisage une solution logicielle dans une perspective multi-agents, le concepteur doit gérer la relation entre :

- le niveau macroscopique du Système Multi-agents (SMA) (la société des agents), où se posent les problèmes d'ingénierie des interactions et d'organisation de la société du SMA afin d'obtenir, du point de vue global du système, les fonctionnalités correspondant aux exigences de l'utilisateur ;
- le niveau microscopique du SMA (les agents en eux-mêmes) où se posent les problèmes d'identification des rôles nécessaires, d'ingénierie des comportements tenant compte des interactions qui se produiront et fournissant les différentes compétences recherchées.

Notre approche suivra les méthodes AALAADIN [Ferber, Gutknecht, 1995] et GAIA [Wooldridge, 1999] qui expriment le souci de chercher à concevoir un SMA comme une organisation humaine, en identifiant les rôles nécessaires au fonctionnement des sociétés d'agents, les relations qui existent entre ces rôles et les interactions systématiques auxquelles ils participent suivant des protocoles institutionnalisés. Nous partirons des fonctionnalités du système exprimées au niveau social pour aboutir au comportement interne des agents. Nous envisageons d'utiliser à plus long terme la plate-forme JADE [Bellifemine et al., 2001], développée par l'université de Parme pour l'implantation du système.

7.1.1 Organisation des sous-sociétés

Dans les travaux du projet CoMMA, lors de l'analyse des sous-sociétés dédiées aux ressources (modèles, documents et pages jaunes), [Gandon, 2002] a constaté la récurrence de trois types de sociétés.

La société hiérarchique (figure 24) distingue deux types d'agents : (1) Les représentants : médiateurs entre leur société et le reste du SMA. Ils sont responsables du traitement des requêtes externes, de leur décomposition, du dialogue avec les exploitants et de l'intégration des résultats intermédiaires pour

fournir une réponse appropriée au commanditaire externe ; (2) Les exploitants : assignés à une ressource locale, ils contribuent autant que possible à la résolution des requêtes qu'ils reçoivent, avec cette ressource locale. Dans cette société, l'information est répartie entre les agents assignés aux ressources. Ceci permet de conserver la répartition initiale des ressources et de distribuer et équilibrer la charge de travail : le travail de recherche locale est réparti entre les exploitants et le travail de fusion est effectué par les représentants. La spécialisation des agents et la distribution des rôles permettent la distribution de la charge de travail mais en contrepartie requièrent un volume plus important d'échanges sur le réseau.

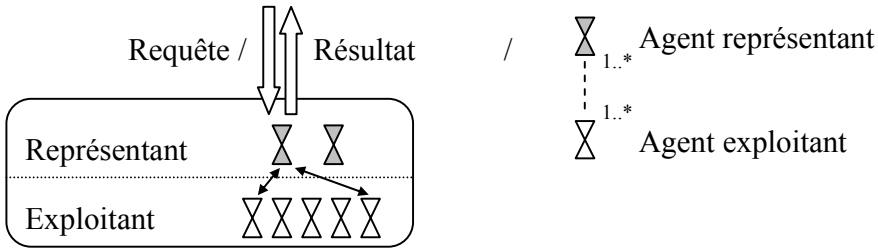


Figure 24 Société hiérarchique

La société égalitaire (figure 25) repose sur des relations égalitaires entre des rôles complètement redondants. N'importe quel agent peut être contacté de l'extérieur de la société pour répondre à une requête concernant le type de ressources auquel sa société est dédiée. Il doit alors coopérer avec ses pairs pour résoudre efficacement la requête.

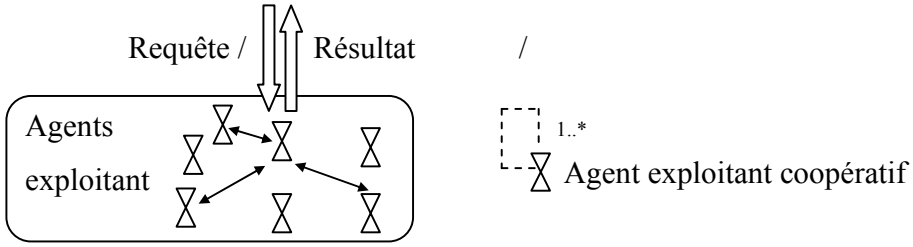


Figure 25 Société égalitaire

Dans cette société, les agents sont uniquement spécialisés par le contenu de la ressource locale à laquelle ils sont assignés. La charge de travail est moins distribuée que dans l'exemple précédent mais la charge réseau peut être diminuée. Il

n'y a qu'un seul type d'agent et il joue les deux rôles précédents. Chaque agent est capable de former une coalition avec d'autres agents pour résoudre les requêtes externes.

La société de duplication (figure 26) est un cas particulier du cas précédent : ni les rôles ni le contenu ne sont distribués. Chaque agent maintient à jour une copie complète de toute l'information et peut résoudre les requêtes seuls. Par conséquent, les seules interactions sociales qui existent concernent les mises à jour du contenu. La charge de travail est bien moins distribuée que dans le cas précédent et le contenu doit être dupliqué auprès de chaque agent de la société ce qui peut être une contrainte inacceptable. En revanche, il n'y a aucune spécialisation, le système est fortement redondant, et par conséquent fortement résistant aux pannes. L'utilisation du réseau est minimale lors du traitement d'une requête. Le seul rôle en commun avec les sociétés précédentes est celui de l'exploitant.

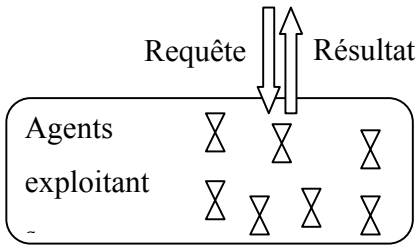


Figure 26 Société de duplication

Selon les tâches à exécuter, la taille des données, et plus généralement selon les contraintes imposées, une sous-société dédiée sera organisée selon l'un ou l'autre des modèles précédents, et les interactions exigeront différents protocoles (requête, question, appel à proposition, enchères, etc.).

7.1.2 Des sociétés en macroscopique

Les fonctionnalités souhaitées pour le système ne se transfèrent pas directement en fonctionnalités d'agents, mais influencent la conception et sont finalement distribuées dans les interactions sociales des agents et l'ensemble des capacités, des rôles et des comportements qui leur sont associés. En considérant les fonctionnalités du système OntoWatch, nous avons identifié cinq sous-sociétés d'agents :

- Une sous-société dédiée à l'ontologie du système,
- Une sous-société dédiée à la recherche sémantique (basée sur des annotations sémantiques existantes),
- Une sous-société dédiée à la recherche sur le Web et à la génération des annotations sémantiques des documents,
- Une sous-société « pages jaunes et blanches » dédiée à l'interconnexion,
- Une sous-société dédiée à l'interface et l'utilisateur.

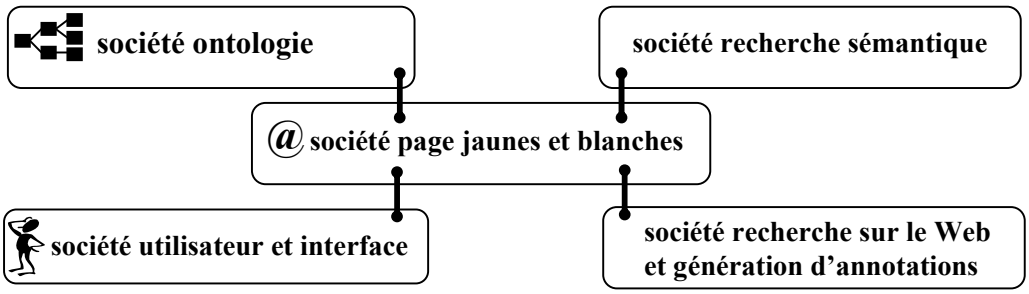


Figure 27 Graphe de voisinage des sous sociétés d'agents

La figure 27 montre le graphe de voisinage entre ces sous-sociétés d'agent Il est clair que la sous-société « pages jaunes et blanches » dédiée à l'interconnexion est l'épine dorsale permettant aux agents d'interagir. Dans les prochaines sections, nous allons raffiner la figure 27 en projetant les nœuds vers des rôles d'agents. Puis les protocoles d'interaction seront introduits, dérivant de leurs rôles sociaux. Les étapes sont :

- Décomposer et identifier la structure convenable pour des différentes sous-sociétés.
- Déterminer les rôles possibles dans chaque société et les protocoles d'interaction entre les agents.

7.1.3 Sous-société dédiée à l'ontologie

Les agents ontologiques [Singh and Huhns, 1999] sont essentiels pour l'interopérabilité. Ils fournissent un contexte commun comme un fond sémantique que les agents peuvent alors utiliser pour relier leurs terminologies individuelles ; ils

fournissent l'accès aux ontologies multiples ou aux différentes parties d'une large ontologie ; enfin ils gèrent l'évolution distribuée de l'ontologie.

Dans le contexte du système d'aide à la veille OntoWatch, l'ontologie est utilisée quand :

- Un veilleur veut exprimer sa cible de veille (concept du domaine d'expertise, type de document, source,…)
- Le système fait la recherche sémantique dans la base d'annotations sémantiques.

Les agents dans la société dédiée à l'ontologie sont les agents fournisseurs des services concernant l'exploitation d'ontologie dans les activités de recherche d'information et de génération des annotations sémantiques. Ces agents vont fournir aux autres agents les services suivants :

- Faire des requêtes sur l'ontologie, en particulier sur les hiérarchies de concepts et de relations de l'ontologie,
- Télécharger et mettre à jour l'ontologie,
- Faire des recherches sur le Web en utilisant les concepts de l'ontologie.

Dans le système de veille OntoWatch, le domaine de veille est varié et peut changer selon le besoin de l'organisation et la taille de l'ontologie de veille peut être très grande avec plusieurs domaines d'application. Dans ce cas, pour réduire la tâche de recherche des concepts nécessaires effectuée par les agents, nous pouvons diviser l'ontologie en plusieurs parties, chacune associée à un agent :

- Une partie dite générique, correspondant à la couche plus haute de l'ontologie, contient des concepts abstraits et très généraux, utilisables pour tous les domaines de veille. Cette partie couvre aussi les concepts concernant les documents et les sources d'information.
- Des parties correspondant à des sous-domaines spécifiques de veille : par exemple, la construction et bâtiment, l'eau et le recyclage, sécurité et incendie.

A partir de cette structuration de l'ontologie, on peut envisager différents rôles d'agents : agent responsable d'un domaine, agent dédié à la partie générique de l'ontologie.

Nous allons également prendre en compte notre algorithme de recherche et d'annotation des documents Web : pour la construction des requêtes systèmes, l'algorithme peut exploiter soit une branche d'ontologie (parcours vertical), soit les concepts à un niveau de profondeur spécifique (parcours horizontal).

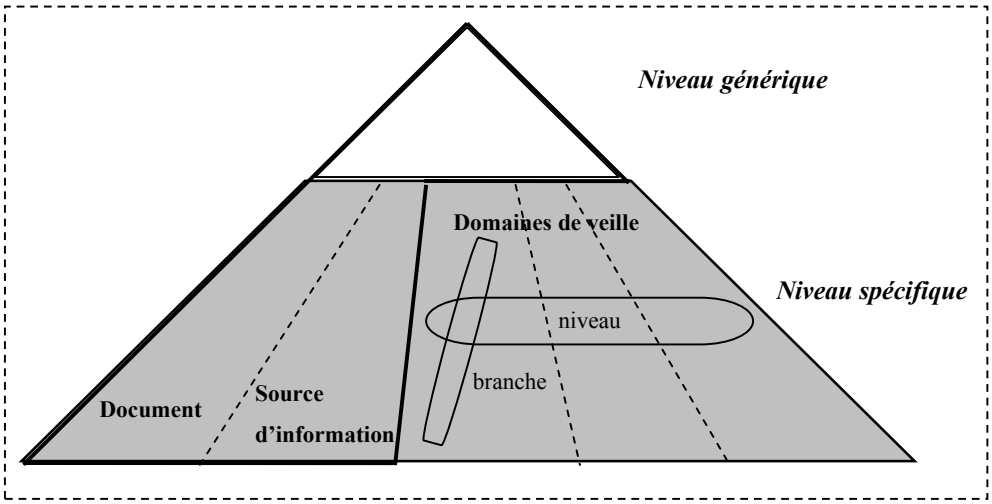


Figure 28 Les différentes parties de l'ontologie gérées par agents

Nous avons donc identifié également les rôles d'agents responsables d'une partie spécifique de l'ontologie comme : agent dédié à une branche et agent dédié à un niveau.

En résumé, dans cette société, interviennent les quatre rôles d'agents suivants :

- Un agent dédié à un niveau de l'ontologie propose aux autres agents clients des services d'interrogation sur les concepts descendants situés à ce niveau de profondeur de l'ontologie.
- Un agent dédié à un sous-domaine spécifique de l'ontologie fournit une vue partielle de l'ontologie, concernant tous les concepts descendants d'un concept spécifique : par exemple, un agent dédié à la construction et au bâtiment, un autre dédié à la sécurité.

- Un agent dédié aux branches de l'ontologie gère toutes les branches d'ontologies ayant un concept spécifique comme racine.
- Un agent dédié au niveau générique « top-level » de l'ontologie propose tous les services de diffusion, de mise à jour et de requêtes sur la partie générique de l'ontologie. En effet, la frontière entre cette partie avec les autres parties de l'ontologie est définie selon le point de vue de certains veilleurs et change avec l'évolution de l'ontologie. Donc cette partie est déterminée selon différents critères ; puis établie dans la phase de configuration des agents.

Ces rôles sont associés aux ressources locales (qui constituent les différentes parties d'une ontologie). Nous avons donc besoin d'un rôle d'agent Médiateur ou Manageur servant de représentant de la société : ce médiateur traduit les requêtes externes, puis distribue les travaux aux agents affectés aux ressources. Nous avons choisi une organisation hiérarchique pour cette société dédiée à l'ontologie.

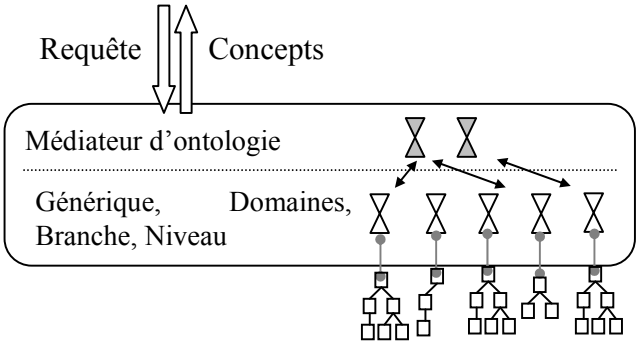


Figure 29 Société dédiée à l'ontologie

7.1.4 Sous-société dédiée à la recherche sémantique

Les agents dédiés à la recherche sémantique sont en charge de l'exploitation des bases d'annotations du système de veille. Grâce à ces annotations, les agents vont avec l'aide du moteur de recherche sémantique, chercher et retrouver les références vers les documents satisfaisant une requête donnée.

Ces agents étant attachés aux annotations, la structure des bases d'annotations dans le système va donc influencer la conception de cette société. Les annotations pouvant être construites et stockées de manière distribuée dans le système

OntoWatch, l'application de la technologie agents pour la gestion des annotations semble raisonnable.

Nous avons prévu que si dans le système, le nombre d'annotations devient de plus en plus énorme dans le temps, le stockage de toutes les annotations dans une seule base peut s'avérer peu efficace ; il peut donc être intéressant de structurer les bases d'annotations. Plusieurs critères peuvent être envisagés pour stocker des annotations dans différentes bases d'annotations, selon trois différents aspects intéressant le veilleur quand il cherche des informations :

- Le sujet d'annotation concerné : on peut ainsi regrouper dans la même base les annotations concernant un sujet spécifique,
- Le type de document annoté : on peut ainsi différencier une base d'annotations sur les journaux, une base d'annotations sur des rapports techniques, etc.
- La source d'information : on peut alors constituer une base d'annotations issues de bases de données en ligne, une base d'annotations pour les instituts de recherche, une autre pour les entreprises.

Nous n'avons pas opté pour la première solution, car elle n'assure pas que les bases d'annotations soient disjointes. Il existe des documents concernant plusieurs domaines d'applications à la fois et donc leurs annotations appartiendraient à plusieurs bases d'annotations à la fois. Cela compliquerait la tâche de gestion effectuée par les agents logiciels. D'autre part, le système OntoWatch est ouvert, donc les domaines gérés par le système évoluent dans le temps et demandent des réglages supplémentaires dans la configuration des agents.

De plus, la nature des sources d'information est en général moins indiquée que la nature des documents recherchés quand les veilleurs effectuent la recherche d'information. Pour cette raison, nous avons décidé d'organiser des bases d'annotations selon le type de document. Pour l'instant nous envisageons cinq bases d'annotations correspondant aux annotations sur : les rapports, les articles, les thèses, les brevets et les autres documents (dont le type n'est pas encore déterminé dans les annotations).

Chaque base d'annotations est gérée par un agent dédié. Chaque agent est dédié à une ressource spécifique. Quand le système transforme une requête de l'utilisateur en une requête système, cette requête va être attribuée à un agent selon son contenu. Comme les ressources locales (base d'annotations) sont clairement distinguées, on peut considérer que les agents dédiés à chaque base d'annotation sont appropriés au rôle des agents dédiés à des ressources et que l'agent médiateur d'annotations joue le rôle d'agent représentatif de la société. Nous avons donc opté pour une organisation hiérarchique pour cette société. Pour cela, nous avons :

- un rôle *Médiateur* d'annotations en charge de gérer des requêtes extérieures, de déterminer la base d'annotations concernée, et puis d'attribuer le travail de recherche à l'agent dédié à cette base,
- un rôle *Archiviste* d'annotations associé à la gestion d'une base d'annotations correspondant au type de document défini. Chaque Archiviste doit offrir des services de recherche sémantique (en utilisant CORESE) sur sa base d'annotations, et des services d'archivage des nouvelles annotations dans sa base.

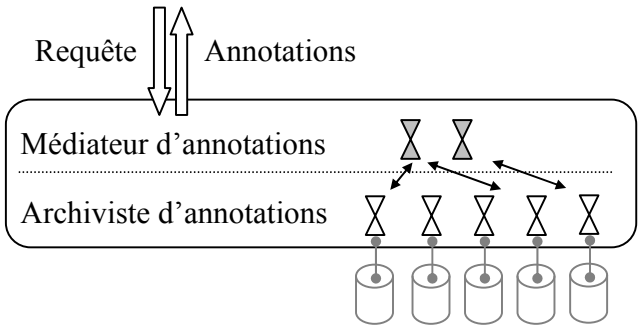


Figure 30 Société dédiée à la recherche sémantique

7.1.5 Sous-société dédiée à la recherche sur le Web et à la génération des annotations sur les documents Web

Les agents de cette société sont responsables d'une tâche importante du système : effectuer la recherche sur le Web externe en utilisant les moteurs de recherche tel que Google et Yahoo!, en étant guidés par l'ontologie pour générer des annotations sémantiques et alimenter des bases d'annotations. Cela signifie que ces agents sont

chargés de distribuer les travaux des algorithmes dans le chapitre précédent. Les rôles d'agents doivent être conçus en analysant ces algorithmes. Rappelons que l'idée principale de tous les algorithmes est de compléter la requête de l'utilisateur en générant plusieurs requêtes système et que le système va chercher des documents sur le Web avec ces requêtes. Ces agents vont exploiter l'ontologie, travailler avec des moteurs de recherche, et finalement générer de nouvelles annotations. Ces agents vont coopérer avec les agents de la société dédiée à la recherche sémantique pour faire la mise à jour des bases d'annotations avec ces nouvelles annotations. Cette société d'agents a besoin d'un rôle d'agent local chargé de s'occuper de toutes les tâches nécessaires pour traiter une requête système : faire la recherche avec Google et extraire des informations pour générer ensuite une annotation pour chaque document réponse. En outre, le rôle d'un médiateur responsable du contact avec l'extérieur est indispensable : ce médiateur devra transformer la requête de l'utilisateur en plusieurs requêtes système et attribuer les tâches aux agents locaux, puis faire la synthèse des résultats retournés par ces agents pour générer des annotations correspondant à la requête originale.

Concernant cette sous-société, seuls les deux premiers types d'organisations sont envisageables :

- Dans une société hiérarchique : nous introduirons un rôle d'agent *Annotateur* en charge de gérer les requêtes posées par l'utilisateur (dans le cas où le résultat de la recherche sémantique dans la base d'annotations n'est pas satisfaisant). Nous aurons aussi un rôle d'agent *Collecteur* responsable de la recherche sur le Web avec une requête système et de l'extraction des informations nécessaires à retourner à l'agent Annotateur afin de générer des annotations.
- Dans une société égalitaire, nous aurions un rôle d'agent *Annotateur* co-opératif combinant les deux rôles précédents.
- Une société de duplication pour la recherche sur le Web et l'annaotation n'est pas réaliste car cela signifierait la duplication d'une image complète de toutes les annotations engendrées pour chaque instance d'un agent de cette

société. Ceci n'est pas concevable dans le cadre d'un nombre important de requêtes soumis au système de veille et d'un grand nombre d'annotations générées correspondant à chaque requête.

Pour le système OntoWatch, nous avons choisi la première solution pour la société dédiée à la recherche et l'annotation des documents Web. Le rôle d'*Annotateur de documents Web* propose ses services aux agents des autres sociétés pour enrichir leurs requêtes de recherche sur le Web, et pour générer leurs annotations. L'*Annotateur* engage les services des Collecteurs de documents Web pour étendre effectivement la recherche et obtenir des annotations :

- A partir de la requête de l'utilisateur, l'Annotateur génère plusieurs requêtes système en utilisant l'ontologie,
- Il distribue les requêtes systèmes aux Collecteurs,
- Il combine les réponses partielles pour générer des annotations finales (comparaison des URL, combinaison des listes de concepts, etc.).

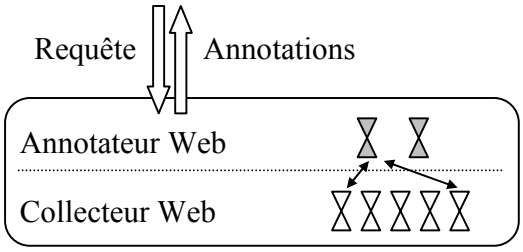


Figure 31 Société dédiée à la recherche sur le Web

Le rôle du Collecteur du Web est d'envoyer une requête système contenant des mots-clés correspondant à des concepts de l'ontologie à Google, puis de traiter la réponse pour obtenir des informations nécessaires pour annoter chaque document.

7.1.6 Sous-société dédiée à l'interconnexion

Les agents de cette sous-société sont responsables de la connexion, par appariement, des autres agents en se basant sur l'expression de leurs besoins et la description de leurs capacités. Chaque agent fournisseur d'un service doit enregistrer sa description auprès d'un agent d'interconnexion pour que les attentes des agents demandeurs puissent y être comparées à cette description. La littérature propose plusieurs types

d'agents d'interconnexion et la terminologie n'est pas consensuelle : la même désignation peut dénoter des rôles différents. Nous adoptons les définitions de [Klusch, 1999] :

- Un agent courtier (broker) identifie les fournisseurs potentiels, leur transmet la requête, récupère les résultats et renvoie une réponse complète au demandeur.
- Un agent apparieur (matchmaker) identifie les fournisseurs potentiels, fournit cette liste de candidats au demandeur et le laisse prendre contact avec eux.

Les agents du système OntoWatch seront implémentés en utilisant la plate-forme JADE [Bellifemine et al., 2001] qui fournit un type d'agent appelé Directory Facilitator et un autre type d'agent appelé AMS (*Agent Management System)*:

- Dans JADE les agents responsables des pages jaunes DFs (Directory Facilitators) sont des apparieurs que l'on peut fédérer en une société égalitaire. Le service des pages jaunes permet de fournir l'adresse d'un agent en fonction d'une ou plusieurs de ses capacités. Selon les spécifications de FIPA, les DFs offrent des identificateurs d'agent appariant la description de service et l'ontologie spécifiée dans un pattern. Ainsi les DFs sont des apparieurs qui identifient les fournisseurs appropriés et renvoient la sélection des candidats au demandeur. Le résultat de l'appariement peut être encore raffiné dans une deuxième étape.
- L'agent AMS gère le service de pages blanches qui donne l'adresse d'un agent en se basant sur son nom.

7.1.7 Sous-société dédiée à l'utilisateur

Les agents de la sous-société dédiée aux utilisateurs sont en charge des aspects d'interface, d'observation et d'aide à l'utilisateur. Ces agents sont en général demandeurs de services. Etant donné qu'ils ne sont pas liés à une ressource comme dans les cas précédents, ces agents ne suivent pas la typologie des sous-sociétés que nous venons d'exposer. Les rôles définis dans cette sous-société et leur distribution dépendent de caractéristiques fonctionnelles supplémentaires du système. Nous nous sommes intéressés en priorité à deux rôles : (1) Le rôle de gestion de

l'interface utilisateur : en charge du dialogue avec l'utilisateur, il doit permettre l'expression de requêtes par l'utilisateur, leur raffinement et la présentation des résultats dans un format approprié. (2) Le rôle de gestion du profil utilisateur : les agents de profil d'utilisateur sont similaires aux agents archivistes mais leurs annotations concernent les utilisateurs. Les profils sont exploités d'abord pour des aspects d'interfaces. Dans ce travail comme nous n'avons pas approfondi ces aspects interfaces, nous avons donc décidé de réutiliser les agents de même fonctionnalité que dans le système CoMMA [Gandon, 2001].

7.1.8 Vue globale des sous-sociétés

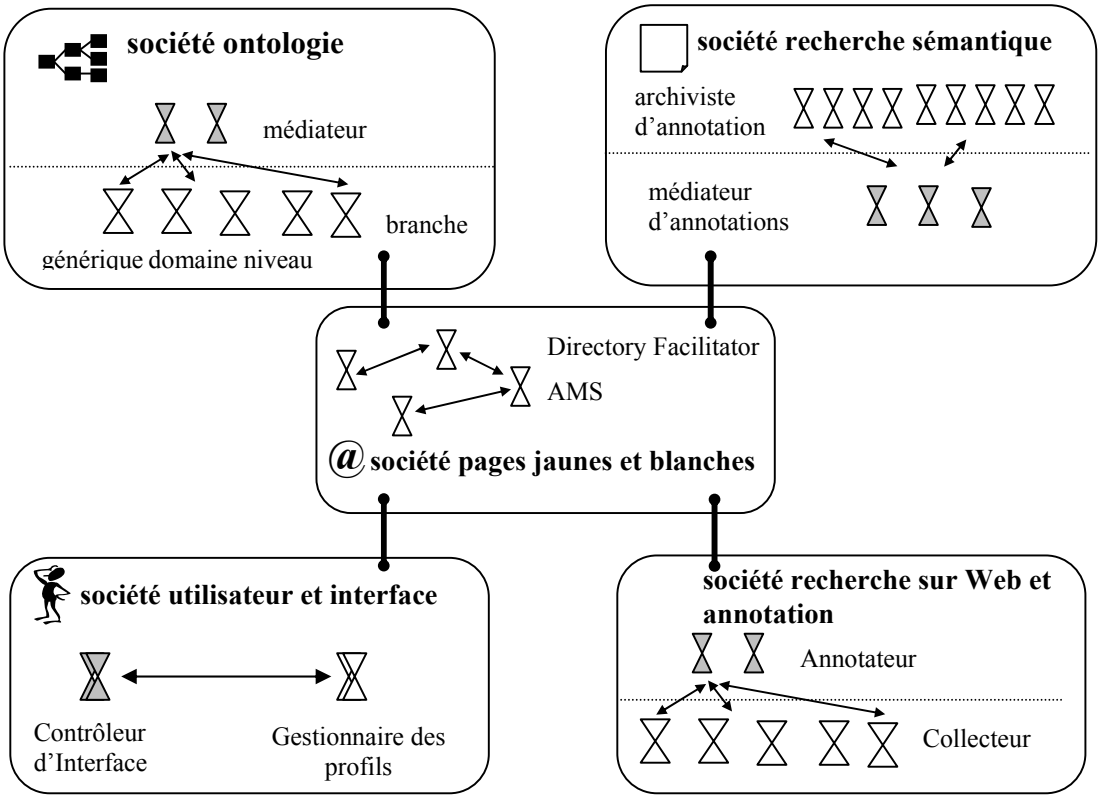


Figure 32 Sous sociétés d'agents et leur organisation interne

La figure 32 montre les sous-sociétés d'agents dans le système OntoWatch avec de nouveaux rôles identifiés. Treize rôles ont été identifiés pour être implémentés dans le système : Médiateur d'ontologie, agent dédié à une Branche d'ontologie, agent dédié à un Niveau d'ontologie, agent dédié à un sous Domaine d'ontologie, agent

dédié à l'ontologie générique, Médiateur d'annotations, Archiviste d'annotations, Apparieur (DF, AMS), Contrôleur d'interface, Gestionnaire des profils, Annotateur des documents, Collecteur des documents.

7.2 Des rôles aux interactions

La deuxième étape de notre approche est de dériver de l'analyse de l'architecture les caractéristiques des rôles, les protocoles de leurs interactions afin de choisir une implantation des comportements de chaque type d'agents.

7.2.1 Les rôles

Un rôle représente la position d'un agent dans une société, les responsabilités et les activités assignées à cette position que les autres agents s'attendent à voir correctement remplies. L'analyse des rôles est à la charnière entre le niveau microscopique des agents et le niveau macroscopique du SMA. La partie précédente a identifié les rôles à l'intérieur des sociétés ; nous étudions maintenant le niveau microscopique en considérant le système du point de vue d'un agent qui joue un rôle. Pour décrire ces rôles, nous adaptons la carte de rôle proposée par E. Kendall [Kendall, 1999]. Les définitions des facettes d'un rôle d'agent sont :

modèle de rôle	nom de rôle
contexte	la société d'agents concernée
responsabilités	services, tâches, buts, obligations, interdictions
collaborateurs	les rôles avec lesquels il interagit
interfaces externes	l'accès aux services externes ou ressources
expertise	ontologie, inférence, connaissances pour la résolution de problème
interactions	protocole, résolution des conflits, connaissances sur la raison d'interagir avec d'autres rôles, permissions
autres	-

7.2.1.1 Médiateur d'Ontologie

Ce rôle est en charge de la résolution des requêtes externes sur une ou plusieurs ontologies du système de veille. Après cette tâche, il attribue les demandes aux agents dédiés à une partie d'ontologie ou à une ontologie spécifique (dans le cas où

système travaille avec de multiples ontologies). Il interagit avec le DF et tous les agents qui ont besoin des services d'ontologie, concrètement les Annotateurs Web, le Collecteur Web, les Archivistes d'annotations, le Contrôleur d'interface et les agents qu'il gère dédiés aux différentes parties de l'ontologie.

Tout d'abord le Médiateur d'ontologie (MO) va s'inscrire auprès du DF pour indiquer ses services offerts.

modèle de rôle	rôle Médiateur d'ontologie
contexte	la société dédiée à l'ontologie
responsabilités	Fournir des services d'interrogation de l'ontologie à l'exterieur. Gérér les agents dédiés aux différentes parties de l'ontologie
collaborateurs	DF, Médiateur d'annotations, Archiviste d'annotations, Annotateur Web, Collecteur Web, Contrôleur d'Interface, Agent dédié aux sous-domaines de l'ontologie
interfaces externes	Interface de manipulation de l'ontologie, agents dédiés aux branches, agents dédiés aux niveaux de l'ontologie, agent dédié à l'ontologie générique
expertise	Gestion et Interrogation de l'ontologie
interactions	FIPA request protocol (FIPA Query-Ref protocol)
autres	-

Il va également accepter l'inscription des agents dédiés aux parties de l'ontologie et devient leur gestionnaire et leur représentant.

Le Contrôleur d'Interface le contacte quand l'utilisateur veut choisir les concepts pour former des requêtes. L'Annotateur Web interagit avec le MO pour interroger l'ontologie afin d'étendre la requête utilisateur. En fonction de la partie de l'ontologie concernant la demande de l'Annotateur Web, le Médiateur d'ontologie va attribuer le travail à un agent local approprié (dédié à l'ontologie générique, dédiés aux sous-domaines, dédiés aux branches, dédié aux niveaux). De la même manière, le Collecteur Web interagit avec le MO pour comparer les termes extraits avec les concepts de l'ontologie. L'interaction entre le Médiateur d'ontologie et le Médiateur d'annotation va être décrite dans la section 7.2.1.10. Les Archivistes

d'annotation devraient interagir avec le MO lorsque ce dernier fait des recherches sémantiques sur les bases auxquelles ils sont associés.

7.2.1.2 Agents dédié aux sous-domaines de l'ontologie

Ce rôle d'agent propose des services d'interrogation sur les grands sous-domaines de l'ontologie. Il reçoit les demandes provenant du Médiateur d'Ontologie et retourne les concepts, relations satisfaisant ces demandes. Une fois créé, il va contacter le DF pour trouver le Médiateur d'ontologie, puis s'inscrire et désinscrire ses services avec le Médiateur d'ontologie. Dès ce moment, il est sous la gestion de ce Médiateur d'ontologie et il ne connaît et interagit qu'avec ce MO.

modèle de rôle	rôle agent dédié aux sous-domaines de l'ontologie
contexte	la société dédiée à l'ontologie
responsabilités	Fournir des services d'interrogation sur des grands sous-domaines d'une ontologie
collaborateurs	DF, Médiateur d'ontologie
interfaces externes	Interface de manipulation de l'ontologie
expertise	Gestion et Interrogation de l'ontologie
interactions	FIPA request protocol (FIPA Query-Ref protocol)
autres	-

7.2.1.3 Agents dédié aux branches de l'ontologie

Ce rôle d'agent propose des services d'interrogation sur les branches de l'ontologie. Il reçoit les demandes provenant du Médiateur d'Ontologie et retourne les concepts et les relations satisfaisant ces requêtes. Ce rôle d'agent est invoqué quand le système utilise des branches d'ontologie pour enrichir une requête de l'utilisateur. Une fois créé, il va contacter le DF pour trouver le Médiateur d'ontologie, puis s'inscrire et désinscrire ses services avec le Médiateur d'ontologie. Il est ensuite sous la gestion de ce Médiateur d'ontologie comme les autres agents locaux dans cette société.

Les deux rôles qui restent : agent dédié aux niveau de l'ontologie et agent dédié à l'ontologie générique sont similaires aux deux rôles décrits précédemment. La

différence n'est que la partie de l'ontologie dont l'agent est responsable. Ces agents sont tous sous la gestion du Médiateur d'Ontologie, et ils n'interagissent qu'avec le DF et le MO.

modèle de rôle	rôle agent dédié aux branches de l'ontologie
contexte	la société dédiée à l'ontologie
responsabilités	Fournir des services d'interrogation sur les branches de l'ontologie.
collaborateurs	DF, Médiateur d'ontologie
interfaces externes	Interface de manipulation de l'ontologie
expertise	Gestion et Interrogation de l'ontologie
interactions	FIPA request protocol (FIPA Query-Ref protocol)
autres	-

7.2.1.4 Contrôleur d'interface

modèle de rôle	rôle Contrôleur d'Interface
contexte	la société dédiée aux Utilisateurs
responsabilités	Manipuler l'interaction directe avec l'utilisateur. Présenter une vue uniforme du système OntoWatch.
collaborateurs	Médiateur d'Annotations, Annotateur Web, DF, Médiateur d'Ontologie, Gestionnaire de profils
interfaces externes	Interface graphique,
expertise	Gestion des interactions avec les utilisateurs
interactions	FIPA request protocol
autres	-

L'agent contrôleur d'interface (CI) est le système frontal, fonctionnant en collaboration étroite avec l'utilisateur. Une caractéristique de l'agent CI est son Interface Graphique utilisateur (GUI), par laquelle les utilisateurs peuvent exploiter le système OntoWatch. Cet agent considère l'utilisateur comme un agent aux yeux du reste de système, car dès que la couche de GUI sera passée, tout l'échange de l'information se produit au moyen de messages de FIPA ACL. Quand le contrôleur d'interface démarre, il utilise les services de Pages jaunes fournis par le DF pour s'inscrire et est informé (grâce au protocole FIPA-Request) de tous les agents nécessaires. Pour les autres interactions, il collabore avec les agents représentants des autres sociétés tel que le Médiateur d'Annotations, l'Annotateur des Documents

Web, le Médiateur d'Ontologie et évidemment le Gestionnaire des profils qui se trouve dans la même société d'agents.

7.2.1.5 Gestionnaire des profils

modèle de rôle	rôle Gestionnaire des profils
contexte	la société dédiée aux Utilisateurs
responsabilités	Gérer le stockage et l'accès aux profils des utilisateurs.
collaborateurs	Contrôlleur d'Interface, DF, Médiateur d'Ontologie, Médiateur d'annotation.
interfaces externes	Interface de manipulation d'annotation RDF
expertise	Gestion des interaction avec les utilisateurs
interactions	FIPA request protocol
autres	-

Le gestionnaire des profils est chargé de gérer et d'exploiter les bases de profils de l'utilisateur dans le système. Le profil d'un utilisateur est effectivement une annotation RDF sur cette personne. Le Gestionnaire des profils interagit avec le Contrôleur d'Interface via le protocole FIPA-Request pour donner son adresse au IC. Il interagit avec le Directory Facilitator pour s'inscrire et désinscrire ses services et chercher les collaborateur. Il contacte le Médiateur d'annotations dans lors d'une demande d'interrogation sur les profils pour lancer la recherche sémantique.

7.2.1.6 Directory Facilitator

Le « Directory Facilitator » (DF) est l'agent responsable de la maintenance du système de Pages Jaunes où les agents peuvent s'enregistrer et enregistrer leurs capacités et auquel ils peuvent envoyer des demandes de recherche reposant sur la description de services pour trouver leurs collaborateurs. Par conséquent, n'importe quel agent du système OntoWatch qui connaît le DF peut accéder à l'adresse de tous les agents enregistrés en fonction des requêtes au sujet des capacités des agents. Ce rôle est fourni et implémenté dans la plate-forme JADE.

modèle de rôle	rôle Directory Facilitator
contexte	La société dédiée à l'interconnexion
responsabilités	Fournir le service Pages Jaunes de FIPA à d'autres agents / Fédérer avec d'autres agents DF.
collaborateurs	Médiateur d'Annotations, Contrôleur d'interface, et tous les autres rôles

interfaces externes	L'interface de demande pour s'inscrire, se désinscrire, modifier, et rechercher dans la base de données des Pages Jaunes
expertise	Gestion des descriptions de services FIPA
interactions	Répondre au protocole FIPA-Request avec le "FIPA Agent Management Content" pour effectuer des opérations : s'inscrire, désinscrire, modifier et chercher. Initialiser le protocole FIPA-Request avec le "FIPA Agent Managent Content" pour se fédérer avec un autre DF
autres	-

7.2.1.7 Annotateur des Documents Web

L'Annotateur des documents Web (AW) utilise l'ontologie pour enrichir la requête de recherche de l'utilisateur, puis lance la recherche sur le Web en engageant des agents Collecteurs et finalement traite les résultats et génère des annotations sémantiques. D'abord l'Annotateur Web interagit avec DF pour s'inscrire et désinscrire ses services et pour chercher les autres collaborateurs qui sont indispensables pour son travail. Il contacte le Médiateur d'Ontologie pour interroger l'ontologie, le Médiateur d'Annotations pour stocker les annotations générées dans le système de veille. Il reçoit des requêtes de l'utilisateur issues du Contrôleur d'Interface, puis il travaille directement avec les Collecteurs Web. L'interaction avec ces agents se fait par l'échange des messages d'ACL selon le protocole FIPA-request.

modèle de rôle	rôle Annotateur Web
contexte	la société dédiée à la recherche et l'annotation sur le Web
responsabilités	Distribuer la recherche sur le Web aux Rassembleurs Web et combiner des résultats
collaborateurs	Médiateur d'Annotations, DF, Collecteur Web, Médiateur d'Ontologie, Contrôleur d'Interface
interfaces externes	Interface qui accepte des requêtes d'utilisateurs pour chercher sur le Web des documents et stocker leurs annotations
expertise	Extension des requêtes avec l'ontologie, génération des annotations
interactions	FIPA request protocol
autres	-

7.2.1.8 Collecteur des documents Web

Le Collecteur Web est un rôle associé aux requêtes générées par le système de veille (AW). Il est en charge de chercher sur le Web avec ces requêtes en utilisant les moteurs de recherche externes (Google, Yahoo!). Il doit ensuite extraire dans les documents trouvés les informations nécessaires pour la génération des annotations. Comme les autres agents, le Collecteur Web interagit d'abord avec DF pour chercher un Annotateur Web, puis s'inscrire et désinscrire ses services avec l'Annotateur Web. Dès ce moment, ils sont sous la gestion de cet Annotateur Web, et seul cet AW peut l'invoquer. Quand il fait l'extraction d'information dans les documents réponses, il interagit avec un Médiateur d'ontologie pour utiliser les vocabulaires de l'ontologie.

modèle de rôle	rôle Collecteur Web
contexte	la société dédiée à la recherche sur le Web
responsabilités	Rechercher sur le Web avec une requête système provenant d'un Annotateur et retourner les informations nécessaires
collaborateurs	DF, Annotateur Web, Médiateur d'Ontologie
interfaces externes	Interface qui accepte des requêtes d'utilisateurs pour chercher sur le Web des documents et stocker leurs annotations
expertise	recherche sur le Web, extraction d'information
interactions	FIPA request protocol
autres	-

7.2.1.9 Archiviste d'annotations

L'Archiviste d'annotations (AA) est chargé de tout ce qui concerne une base d'annotations dans le système OntoWatch, chaque base est caractérisée par le type de documents dont elle va stocker les annotations. Une fois reçue une requête attribuée par l'agent Médiateur d'annotation, l'Archiviste d'annotations recherche dans la base à laquelle il est associé pour trouver des annotations satisfaisant cette requête. Il est responsable aussi du stockage de toute nouvelle annotation dans sa base (quand le type des documents annotés est le même que le type de document dont il est responsable). D'abord AA interagit avec un DF pour trouver un

Médiateur d'annotations, puis s'inscrire et désinscrire ses services avec le Médiateur d'annotations. Dès ce moment, il est sous la gestion de ce Médiateur d'annotations. Quand lors de son travail, il a besoin des services concernant l'ontologie, il interagit avec le Médiateur d'ontologie. Les messages sont échangés via les protocoles FIPA-Request, et FIPA-Query Ref.

modèle de rôle	rôle Archiviste d'annotations
contexte	la société dédiée à la recherche sémantique
responsabilités	Gestion des bases d'annotation correspondant aux différents types de document. Faire la recherche et l'archivage des annotations
collaborateurs	DF, Médiateur d'Ontologie, Médiateur d'annotations
interfaces externes	Interface qui accepte des requêtes d'utilisateurs pour chercher les documents gérés par le système
expertise	exploitation et archivage des annotations
interactions	FIPA request protocol (FIPA Query-Ref protocol)
autres	Est également une partie de rôle Gestionnaire de profils

7.2.1.10 Médiateur d'annotations

Ce rôle est responsable de l'analyse d'une demande de recherche/d'ajout d'annotations au système, puis de la sélection d'un Archiviste d'annotation approprié (en fonction du type de document correspondant à cet AA) pour effectuer le travail.

Tout d'abord le Médiateur d'annotations (MA) va s'inscrire auprès du DF pour indiquer ses services offerts. Il va également accepter l'inscription des Archivistes d'annotations et devient leur gestionnaire et leur représentant. Il interagit avec le Contrôleur d'Interface qui lui soumet des requêtes de l'utilisateur, puis il attribue la recherche sémantique aux Archivistes d'annotation en fonction du type de document concernant la requête. Pour effectuer des requêtes sémantiques où l'ontologie joue un rôle indispensable, il doit interagit avec un Médiateur d'Ontologie. Il collabore avec l'Annotateur Web quand celui-ci veut stocker une annotation générée.

modèle de rôle	rôle Médiateur d'annotations
Contexte	la société dédiée à la recherche sémantique

responsabilités	Fournir l'accès à la société. Résoudre la requête sémantique et l'attribuer à un Archiviste approprié, retourner les résultats aux agents à l'extérieur
collaborateurs	DF, Médiateur d'Ontologie, Archiviste d'annotations, Annotateur Web, Contrôleur d'Interface
interfaces externes	Interface qui accepte des requêtes d'utilisateurs pour chercher les documents gérés par le système
expertise	Gestion des requêtes sémantiques
interactions	FIPA request protocol (FIPA Query-Ref protocol)
autres	-

7.2.2 Interactions sociales

Pour répondre aux fonctionnalités globales du système, les agents doivent pouvoir communiquer les uns avec les autres pour déléguer des tâches, échanger des informations, coopérer. L'identification de sous-sociétés est suivie par l'identification de rôles, mais en parallèle avec la spécification des rôles, les interactions sociales sont indiquées dans la facette d'interactions de la carte de rôle. Après la spécification des rôles, nous nous intéressons donc aux interactions. L'interaction entre agents est plus complexe qu'un simple envoi d'un message isolé. Le modèle d'une conversation doit être spécifié par des protocoles que les agents doivent respecter pour que le SMA fonctionne effectivement. Les protocoles sont des codes de conduite sociale pour l'interaction ; ils décrivent des procédures standards régulant l'échange d'informations entre agents en institutionnalisant des modèles de communication pouvant survenir entre des rôles identifiés. La spécification d'un protocole part d'un graphe d'accointances au niveau des rôles, qui représente par un graphe orienté les voies de communication existant entre les agents jouant ces rôles. Une voie non orientée dénote que les deux agents jouant les rôles se connaissent l'un l'autre.

L'exemple dans la figure 33 montre le graphe d'accointance entre les quatre rôles : Contrôleur d'Interface (CI), Directory Facilitator (DF), Annotateur Web (AW), Collecteur Web (CW) avant et après une veille effectuée à l'exterieur du système (sur le Web).

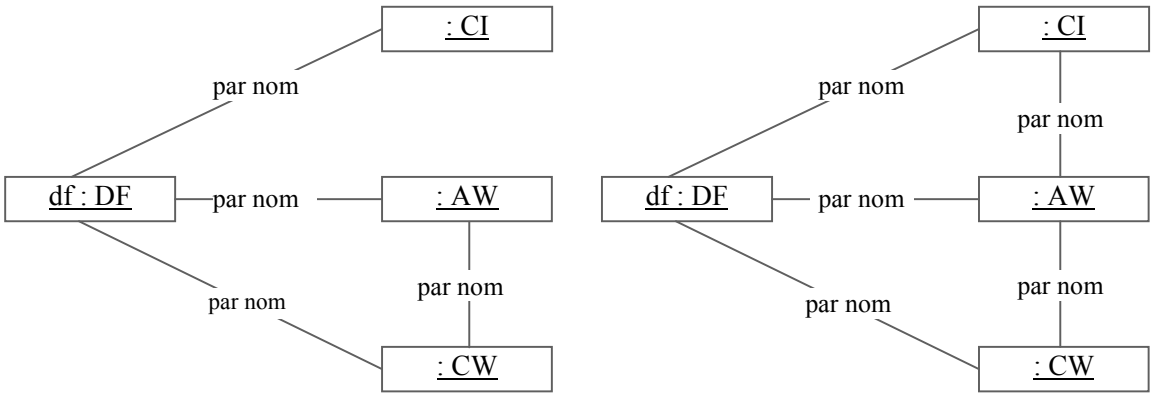


Figure 33 Accointance avant et après une demande de veille sur le Web

Nous voyons sur ces diagrammes qu'au début, le DF connaît les autres agents par son nom et l'Annotateur Web connaît le Collecteur Web car lorsqu'un Collecteur est créé, il doit demander au DF de trouver un Annotateur Web, puis s'inscrire avec cet agent représentant de la société. Mais les interactions concernant la recherche à l'extérieur exigent les accointances entre le Contrôleur d'Interface et l'Annotateur Web.

A partir de ce graphe, on spécifie la séquence possible des messages échangés pendant l'interaction. Le graphe d'accointances se déduit à la fois de l'analyse organisationnelle précédente et des *use cases* du système, décrits à partir des scénarios d'application proposés dans le cahier des charges. Le graphe d'accointances et les messages sont décrits dans des diagrammes de protocoles, une restriction des diagrammes de collaboration UML.

La figure 34 va illustrer par exemple, une sous-partie d'un diagramme d'interaction documentant les échanges ayant lieu lors d'une recherche sur le Web lancée par le système.

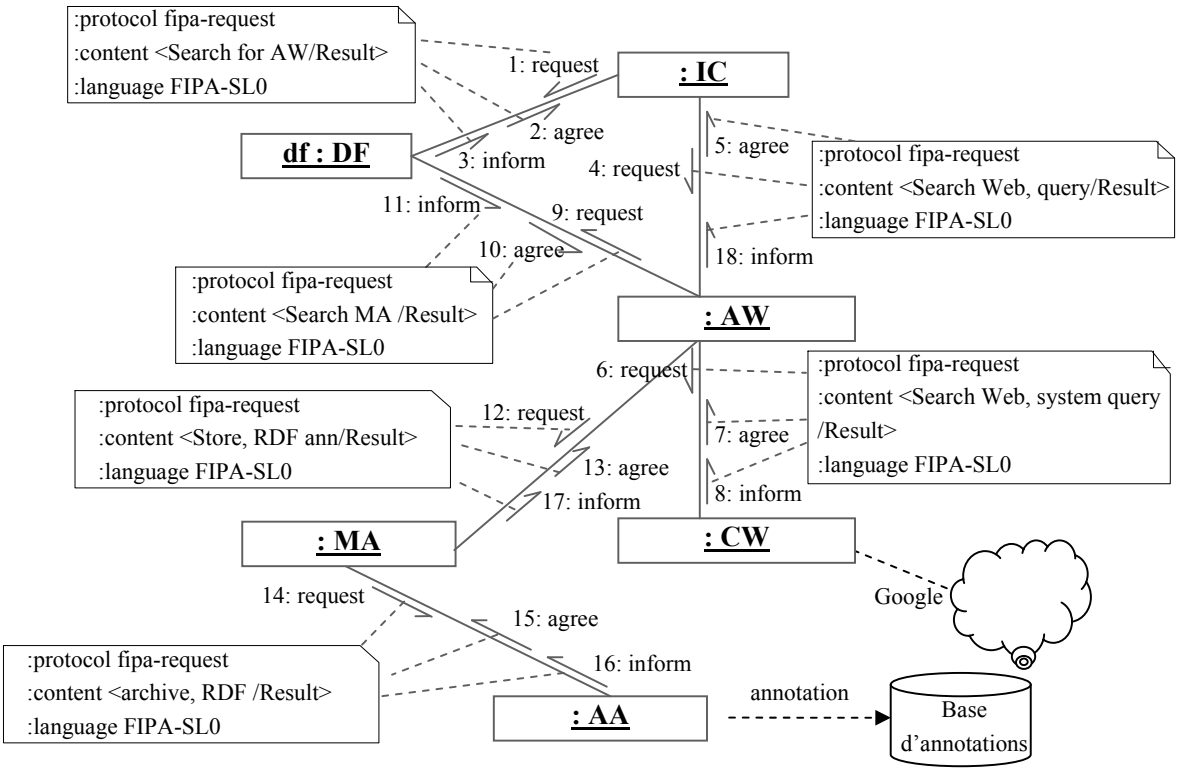


Figure 34 Diagramme d'interactions pour la recherche sur le Web

Le scénario d'interaction est résumé comme suit :

- Le veilleur veut faire une veille sur un sujet et le système décide de faire une recherche à l'extérieur puisque les annotations existantes ne répondent pas au besoin.
- Tout d'abord, le Contrôleur d'Interface (CI) recherche un Annotateur Web (AW) en interrogeant l'agent pages jaunes DF pour obtenir l'adresse d'un AW.
- Puis, CI continue à transmettre une demande de faire une recherche sur le Web avec la requête de l'utilisateur.
- AW est le médiateur dans la société dédiée à la recherche sur Web, il gère les Collecteurs Web (CW). Pour cette raison, AW peut les contacter directement. AW attribue les travaux aux Collecteurs Web et après avoir obtenu les résultats, il génère des annotations RDF.
- Ensuite, AW va demander à DF l'adresse d'un Médiateur d'Annotations (MA) afin de stocker ces annotations.

- Pour répondre à la demande de l'Annotateur Web, le Médiateur d'Annotations va finalement engager des AA pour s'occuper de cette tâche et AW transmet les résultats à CI.

Dans le scénario ci-dessus, nous montrons les interactions entre les agents de différentes sociétés selon l'ordre des principales tâches effectuées par le système. La complexité ne nous permet pas de tous les mettre dans une figure. Dans l'exemple qui suit, nous allons expliquer en détail toutes les interactions entre les agents dans trois sociétés : celle dédiée à la recherche sur l'Internet, celle dédiée à la recherche sémantique et celle dédiée à l'ontologie lors de la tâche de recherche et génération des annotations.

- L'Annotateur Web reçoit une demande de faire une recherche sur le Web avec la requête de l'utilisateur concernant par exemple les deux concepts «recyclage de l'eau» et « bâtiment ».
- Avant d'attribuer les travaux de recherche aux Collecteurs Web, l'Annotateur Web doit utiliser l'ontologie pour générer des requêtes systèmes. Dans cette tâche, à plusieurs moments différents, AW a besoin d'interroger les concepts d'un sous-domaine, des branches ou à partir d'un certain niveau de l'ontologie. Il va demander à DF de trouver un Médiateur d'Ontologie (MO) et envoie au MO la requête indiquant la partie d'ontologie concernée.
- MO analyse la demande du AW et attribue les travaux aux agents dédiés à la partie d'ontologie correspondant. Apres avoir reçu et traité les résultats, MO les retourne à AW pour qu'il puisse continuer sa mission.
- Quand les CW extraient des informations à partir des documents réponse du moteur de recherche, il veut comparer les mots-clés trouvés dans chaque document avec les concepts dans l'ontologie. Il doit contacter DF pour trouver un MO et l'interaction se passe de manière complètement similaire.

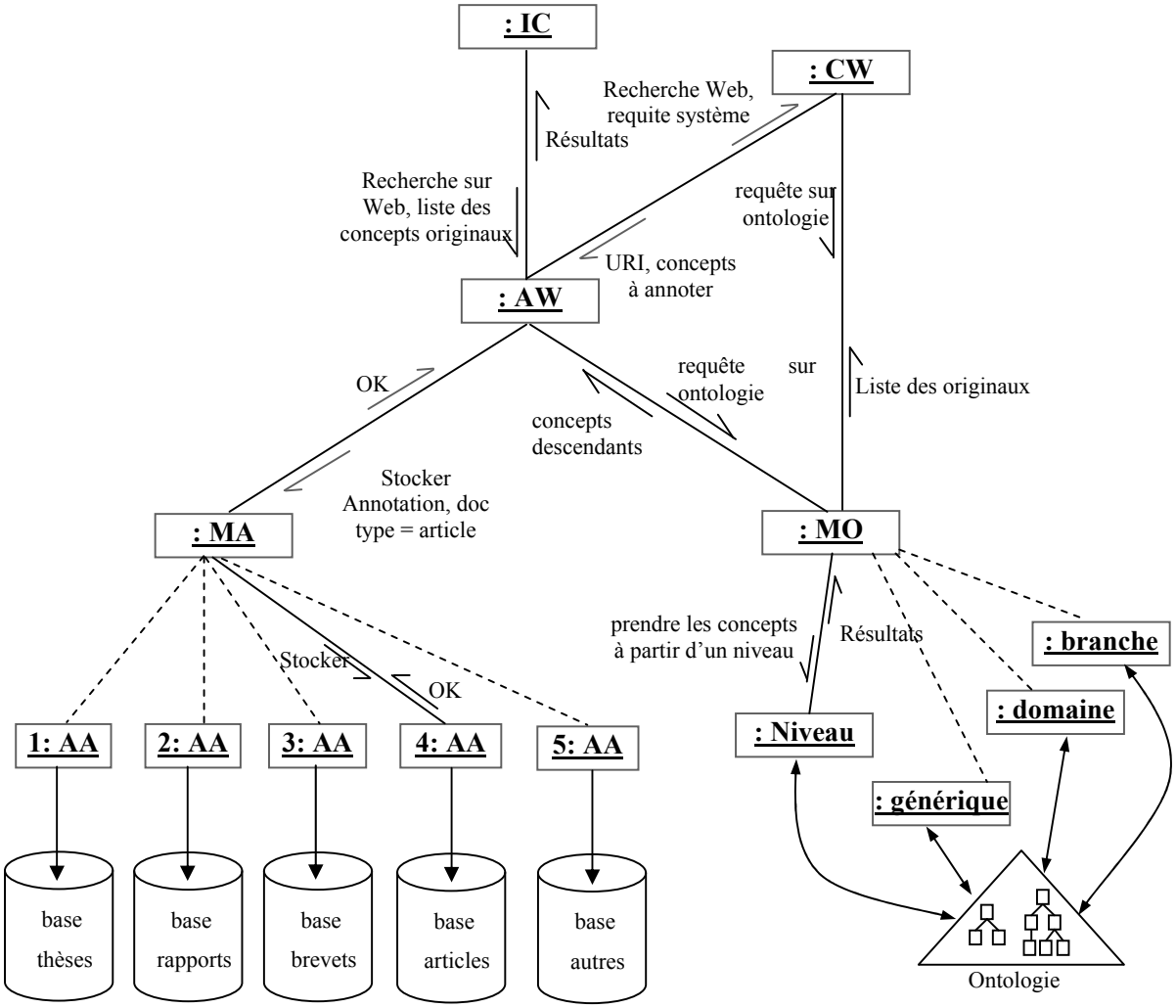


Figure 35 Interactions détaillées entre agents sur l'utilisation de l'ontologie et sur le stockage des annotations

- Après avoir généré une annotation RDF, l'Annotateur Web (via DF) trouve un Médiateur d'annotation pour envoyer une demande de stocker cette annotation, indiquant aussi le type de document annoté. Le type de document est détecté par le AW ; il peut s'agir d'une valeur parmi les suivantes: rapport, thèse, brevet, article et autre.
- Selon le type indiqué dans la requête du AW, Le Médiateur d'Annotations va choisir un Archiviste d'Annotations approprié pour stocker cette annotation.

La figure 35 illustre ces interactions sauf l'interaction entre les agents avec DF pour trouver des agents dont ils ont besoin, pour la simplicité.

A partir de la description des rôles et des interactions, nous pouvons proposer et implanter des types d'agents remplissant un ou plusieurs rôles. Le comportement d'un type d'agent combine les différents comportements implantés pour remplir les activités correspondant aux rôles qui lui sont assignés. Le comportement est fixé par les choix d'implantation déterminant les actions et réactions de l'agent ; ces choix sont libres dans la limite du respect des contraintes imposées par les rôles et les protocoles d'interaction.

7.3 Conclusion

La conception d'architecture démarre du niveau le plus élevé de l'abstraction (cad., la société) et par des raffinements successifs (sous-sociétés), il descend au point où des rôles et des interactions d'agents peuvent être identifiés. Elle se déroule selon les étapes suivantes :

- En étudiant les besoins fonctionnels pour le système au niveau social, nous avons identifié des sous-sociétés d'agents dédiées pour résoudre les différentes facettes de ces fonctionnalités générales.
- Pour chacune de ces sous-sociétés, nous avons identifié un ensemble des types d'organisations possibles pour eux. A partir de l'analyse de l'organisation interne, nous avons identifié des rôles d'agents, qui sont les différentes positions qu'un agent pourrait occuper dans une société et les responsabilités et les activités assignées à cette position et attendues par d'autres pour être accomplies.
- En parallèle avec les descriptions de rôle, nous avons identifié des interactions entre les agents. Les interactions sont précisées avec des protocoles que les agents doivent respecter pour que le système multi-agents fonctionne correctement.

- Finalement, à partir des descriptions de rôle et d'interaction, les développeurs peuvent proposer des types d'agents qui accomplissent un ou plusieurs rôles.

8 Evaluation

Dans ce chapitre, nous présentons les résultats de la phase d'implémentation des algorithmes de la recherche et de la génération des annotations en utilisant l'ontologie. L'intérêt du système de veille est prouvé seulement s'il permet d'obtenir des informations plus pertinentes et plus complètes. Pour cette raison, ces résultats sont vraiment importants. Il est cependant difficile de calculer le taux de rappel et le taux de précision pour des algorithmes qui travaillent sur un immense espace de données comme le Web. Nous nous sommes donc focalisé sur les documents trouvés et annotés par les algorithmes. Nous présenterons tout d'abord notre protocole d'évaluation. Puis, nous analyserons les résultats obtenus.

8.1 Les difficultés de l'évaluation

Plusieurs aspects de l'évaluation sont à prendre en compte pour un système d'information comme OntoWatch. Nous nous sommes focalisé ici sur l'aspect le plus important pour la veille : la recherche d'information, car ce sont les résultats de cette recherche d'information qu'exploiteront les veilleurs. Dans le système Ontowatch, la recherche d'information supportant la veille est effectuée dans deux volets différents :

- La recherche sémantique sur les bases d'annotations RDF,
- La recherche sur l'Internet avec l'aide de l'ontologie.

La recherche sémantique repose sur le moteur Corese, qui est déjà développé et évalué. Cependant, comme la recherche sémantique repose sur l'inférence sur des annotations, la qualité des annotations joue un rôle important.

Pour cette raison, l'évaluation de la recherche sur l'Internet devient primordiale car cette tâche concerne non seulement les réponses directes fournies aux utilisateurs mais aussi les annotations générées par le système et stockées pour être retrouvées lors des recherches sémantique ultérieures.

Normalement, pour évaluer une méthode de recherche, on doit préparer un ensemble de données à tester. Dans le cas d'une recherche de documents, il s'agit d'un corpus comprenant des centaines voire des milliers de documents. Etant donné un algorithme de recherche lancé sur ce corpus, on calcule le nombre de documents trouvés par l'algorithme. Les taux de précision et de rappel sont calculés à partir des résultats selon les formules suivantes :

$$\text{Précision} = \frac{\text{Nombre de documents trouvés corrects}}{\text{Nombre de documents trouvés}}$$

$$\text{Rappel} = \frac{\text{Nombre de documents trouvés corrects}}{\text{Nombre de documents corrects existants}}$$

Pourtant, dans le cas de notre système, un tel calcul est presque impossible. La recherche est menée sur Internet qui contient des milliards de pages Web. Ce

nombre immense de documents dépasse la capacité humaine à les lire tous et à évaluer ceux qui sont corrects ou non pour obtenir le nombre de documents corrects qui auraient dû être trouvés. Comme les algorithmes de OntoWatch utilisent le moteur de recherche Google pour fonctionner, leur performance dans la recherche dépend également de l'efficacité de Google. Nous pouvons seulement évaluer l'efficacité apportée par ces algorithmes grâce à l'utilisation de l'ontologie dans la recherche. C'est ce que nous allons détailler dans la section suivante.

8.2 Le processus de validation

Le but de cette validation est d'obtenir des résultats précis permettant de comparer l'efficacité de la tâche de recherche d'information des veilleurs, avec et sans l'aide du système OntoWatch. Nous allons comparer les résultats obtenus par les recherches et évaluer aussi les annotations générées.

Le protocole d'évaluation d'un algorithme du système OntoWatch est le suivant :

- On définit un sujet de veille.
- On définit également un seuil Max de résultats à examiner.
- Un veilleur (par exemple, dans notre cas, une documentaliste du CSTB) va choisir les concepts dans l'ontologie O'Watch pour exprimer ce sujet.
- Le veilleur va lancer directement la recherche sur Google avec les mots-clés correspondant aux concepts choisis (Recherche Manuelle).
- Le veilleur analyse ensuite les Max premiers résultats de Google pour déterminer lesquels sont des documents satisfaisants par rapport à son objectif de veille et annoter leurs contenus.
- La recherche automatique (Recherche Automatique) par l'algorithme à évaluer est lancée avec les mêmes mots-clés que la recherche manuelle précédente du veilleur.
- L'algorithme renvoie ses résultats, à savoir des documents et leurs annotations générées automatiquement.
- Le veilleur va évaluer les Max premiers résultats retournés par l'algorithme pour décider quels sont les documents pertinents parmi ceux qui ont été annotés par l'algorithme.

Soit M le nombre de bons documents (parmi les Max premiers évalués par les veilleurs) obtenus par la recherche manuelle, soit A le nombre des Max premiers documents trouvés par la recherche automatique, qui sont évalués comme pertinents par les veilleurs.

Soit Doc_M l'ensemble des documents trouvés par la recherche manuelle et Doc_A l'ensemble des documents trouvés par la recherche automatique.

Soit $Doc = Doc_A \cup Doc_M$

Soit $Pertinents_M$ l'ensemble des documents corrects trouvés par la recherche manuelle.

Soit $Pertinents_A$ l'ensemble des documents corrects trouvés par la recherche automatique.

Alors, $M = |Pertinents_M|$ et $A = |Pertinents_A|$

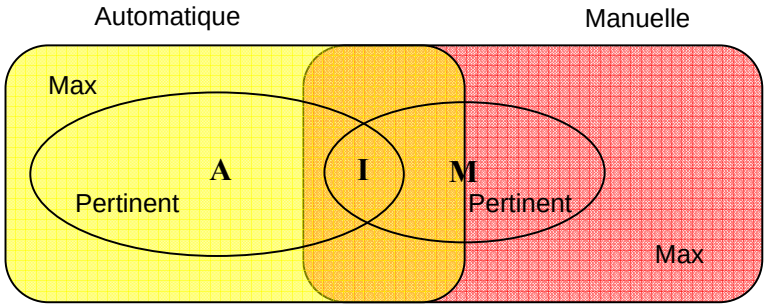


Figure 36 Les mesures pour l'évaluation la recherche automatique et manuelle

On appelle I (Intersection) le nombre de documents pertinents qui existent en même temps dans les résultats de la recherche automatique et de la recherche manuelle.

La comparaison des résultats entre ces deux recherches, l'une manuelle par un humain, l'autre automatique par le système, nous permet de déterminer les mesures suivantes :

$$\text{Nouveauté}_{A/M} = \frac{A - I}{A} \quad \text{Nouveauté}_{M/A} = \frac{M - I}{M}$$

$$\text{Couverture}_{A/M} = \frac{I}{M} \quad \text{Couverture}_{M/A} = \frac{I}{A}$$

$$\text{Pertinence}_{A} = \frac{A}{Max} \quad \text{Pertinence}_{M} = \frac{M}{Max}$$

Le taux de couverture automatique/manuelle donne le pourcentage des documents trouvés par une recherche guidée par l'ontologie dans les résultats de la recherche manuelle normale du veilleur. Réciproquement, le taux de $\text{Couverture}_{M/A}$ indique le pourcentage des documents trouvés par une recherche manuelle normale dans les résultats de recherche guidée par l'ontologie.

La $\text{Nouveauté}_{A/M}$ indique le pourcentage de nouveaux documents pertinents obtenus par le système, qui n'avaient pas été atteints dans les Max premiers obtenus par la recherche manuelle normale.

Par contre, la $\text{Nouveauté}_{M/A}$ est le pourcentage de nouveaux documents pertinents obtenus par la recherche manuelle normale, qui n'avaient pas été atteints dans les Max premiers obtenus par la recherche automatique du système.

Les deux taux de pertinence indiquent les nombres de documents étant évalués par les veilleurs comme pertinents dans les résultats retournés par le système.

Dans ce qui suit, nous présentons le jeu de tests ainsi que les valeurs des mesures d'évaluation présentées ci-dessus.

8.3 Résultats de l'évaluation

Pour mener l'évaluation selon le protocole précédent, nous avons interagi avec une documentaliste du CSTB (qui fait de la veille pour le CSTB) et des experts du CSTB. Nous nous sommes appuyés sur les requêtes mises en œuvre dans le processus de veille sur un domaine particulier, à savoir le recyclage de l'eau. Nous avons testé sur ces requêtes les algorithmes utilisant l'ontologie O'Watch pour chercher et annoter des documents Web.

Soulignons tout d'abord les caractères importants d'une recherche menée par un veilleur du CSTB :

Evaluation

- Pour chaque recherche, Google peut retourner des milliers des documents Web dans les résultats, et le veilleur évidemment ne les analyse pas tous. Il va analyser au maximum les 5 premières pages présentant les résultats, c'est-à-dire 50 documents.
- Pour chaque document, le veilleur va lire le contenu du document, et s'il le juge pertinent par rapport à ce qu'il visait pour sa veille, le veilleur va noter le titre du document, son URL, la date du document et enfin les sujets concernés. Selon le document, le veilleur peut ajouter d'autres informations dans son annotation par exemple la source du document, la nature du document, etc.

En fonction de ces caractéristiques, nous avons configuré l'algorithme pour qu'il analyse seulement les 50 premiers documents retournés par Google pour chaque requête. Nous avons également amélioré l'algorithme pour qu'il essaie d'extraire les mêmes informations que celles apparaissant dans les annotations du veilleur.

La première requête sur la «Réutilisation des eaux grises et pluviales» a été exprimée via les concepts : « water reuse », « rainwater » et « grey water ». Il faut souligner que l'utilisation de l'ontologie permet au système d'effectuer des recherches multilingues, dans le cas de l'ontologie O'Watch en anglais et en français. La définition de chaque concept spécifie ses labels à la fois en anglais et en français, ce qui permet aux algorithmes de prendre la langue comme un paramètre de recherche. Par exemple, le concept « water reuse » est défini dans l'ontologie comme suit :

```
<rdfs:Class rdf:ID="WaterReuseTopic">
 <rdfs:subClassOf rdf:resource="#WaterEconomizingTopic"/>
 <rdfs:label xml:lang="en">water reuse</rdfs:label>
 <rdfs:label xml:lang="en">water recycling</rdfs:label>
 <rdfs:label xml:lang="fr">réutilisation de l'eau</rdfs:label>
 <rdfs:label xml:lang="fr">recyclage de l'eau</rdfs:label>
</rdfs:Class>
```

Les mots-clés en français de ces trois concepts sont : « réutilisation de l'eau » « l'eau pluviale » et « l'eau grise ». Dans ce domaine de veille, nous nous intéressons à la recherche en anglais selon la préférence du veilleur.

Pour cette requête, nous avons obtenu les résultats :

M	Max	A	I
19	50	43	18

$Pertinent_M$	$Pertinent_A$	$Couverture_{M/A}$	$Couverture_{A/M}$	$Nouveauté_{M/A}$	$Nouveauté_{A/M}$
38%	86%	41,86%	94,7%	5,26%	58,13%

Grâce à l'aide de l'ontologie, OntoWatch genère des requêtes systèmes comprenant des concepts descendants de ces trois concepts dans l'ontologie, et la recherche avec ces requêtes enrichies retourne 25 nouveaux documents par rapport aux 19 documents pertinents trouvés par le veilleur dans la recherche manuelle normale, ce qui constitue un résultat intéressant. En analysant le contenu de ces nouveaux documents, nous avons compris la raison pour laquelle ces documents sont apparus dans les résultats de la recherche guidée par l'ontologie. Ils contiennent tous des mots-clés qui n'apparaissent pas dans la requête de l'utilisateur, mais correspondent à des concepts descendants. Donc, avec la recherche manuelle normale, Google ne peut pas les retrouver. Par exemple, le nouveau document intitulé «*Sustainable Water Challenge Rockdale City Council Rainwater Reuse »*, est trouvé par le système avec le mot-clé stormwater, qui correspond au concept « stormwater» - un concept descendant du concept « rainwater ». Deux autres ont été trouvés avec ce concept descendant : l'un est la description d'un projet sur la réutilisation des eaux pluviales, Rockdare Council City Project, l'autre est un rapport de la ville de Toronto.

L'algorithme du système OntoWatch permet aussi de détecter des documents contenant des synonymes des mots-clés utilisés dans la requête de l'utilisateur. Dans ce cas, le système peut retrouver les documents concernant «gray water» au lieu de « grey water ».

Le nombre des annotations qui ont été évaluées comme pertinentes est 43 sur 50. Bien que la requête système couvre la requête utilisateur, il y a un document trouvé manuellement qui n'apparaît pas dans les résultats de la recherche automatique. En

fait ce document appartient aussi aux résultats de Google, mais il ne se trouve pas dans les 50 premiers documents que l'algorithme a traités.

Pour la deuxième requête concernant les concepts : « climate change » « building » « adaptation » et « extreme event » (les mots-clés équivalents en français sont : « changement climatique », « bâtiment », « adaptation », « événement extrême»), les résultats sont montrés dans les tables suivantes :

M	Max	A	I
11	30	18	10

$Pertinent_M$	$Pertinent_A$	$Couverture_{M/A}$	$Couverture_{A/M}$	$Nouveauté_{M/A}$	$Nouveauté_{A/M}$
36,66%	60%	55,55%	90,9%	9,09%	44,44%

Nous pouvons noter une différence remarquable avec le résultat du test précédent, qui s'explique par le nombre des concepts descendants des concepts utilisateurs. Pour cette deuxième requête, certains concepts n'ont pas de concepts descendants dans l'ontologie, par exemple le concept « adaptation » et le concept « climate change ».

Comme toutes les requêtes systèmes sont générées à partir des concepts dans l'ontologie, les documents sont annotés correctement avec les concepts correspondants.

Après analyse de l'expérimentation sur les algorithmes, notamment l'algorithme basé sur une distribution équilibrée des concepts utilisateurs, nous avons également noté que les résultats de nos algorithmes guidés par l'ontologie pour rechercher et annoter des documents Web dépendent de plusieurs facteurs : les caractéristiques de l'ontologie, le nombre de termes utilisés dans la requête de l'utilisateur, le niveau de précision quand les veilleurs utilisent ces termes.

8.3.1 "Ontologie profonde" contre "Ontologie plate"

La structure de l'ontologie influence fortement les résultats de chaque algorithme. Nous appelons "ontologie profonde" une ontologie dans laquelle la majorité des

concepts généraux ont plusieurs niveaux de spécialisation, et inversement, dans une "ontologie plate", un concept général peut avoir beaucoup de sous-concepts directs mais le nombre de niveaux de spécialisation est peu élevé.

Une ontologie plate ne semble pas convenir aux deux premiers algorithmes, qui reposent sur le contenu des branches de l'ontologie pour générer la requête. Dans une ontologie plate, les domaines d'expertise portés par ces branches ne sont souvent pas connexes, par conséquence il y aura peu d'intersections entre les documents trouvés par des requêtes différentes. La probabilité d'obtenir des documents évoquant tous les sujets indiqués dans la requête de l'utilisateur n'est pas grande. Cependant, avec "une ontologie profonde", ces algorithmes fonctionnent bien et permettent de trouver des documents plus spécialisés liés aux sujets auxquels le veilleur est intéressé.

8.3.2 Nombre de concepts dans la requête de l'utilisateur.

Le cas idéal serait que dans chaque requête système générée, le système remplace chaque concept utilisateur par un seul concept de ses concepts descendants. Dans ce cas-là, nous serions assurés de ne manquer aucun concept relatif au sujet surveillé par le veilleur. Mais le problème, lors de l'utilisation de Google comme moteur de recherche, est que le nombre de réponses à une requête simple est énorme. Un petit nombre de concepts dans la requête conduit à une masse énorme de documents trouvés par Google. En outre si le nombre de requêtes générées est trop grand, la limite de Google sera dépassée, ou la charge de la tâche d'élimination des redondances sera augmentée. Réciproquement, quand le nombre de concepts indiqués dans la requête est trop grand, l'algorithme risque d'obtenir moins de documents.

Tous les algorithmes ci-dessus ont une stratégie prédéfinie de sélection de concept, indépendante du nombre de concepts initiaux ; donc aucun algorithme n'assure la meilleure performance dans tous les cas possibles. En fait, quand un veilleur recherche de l'information sur le Web, son expérience le guide pour décider des meilleurs concepts à utiliser dans sa requête initiale.

8.3.3 Le degré de précision du choix des concepts initiaux dans la requête de l'utilisateur

Ce facteur dépend du degré des connaissances du veilleur sur le domaine auquel il est intéressé. Si le veilleur connaît assez bien le sujet à surveiller, il utilisera immédiatement des concepts précis dans sa requête. Ces concepts précis résident à un niveau profond dans l'ontologie. Réciproquement, si le veilleur ne connaît pas bien le domaine à surveiller, il aura tendance à utiliser en premier les concepts généraux correspondant à un faible niveau de profondeur dans l'ontologie. Comme le principe de notre nouvel algorithme est d'assurer une distribution équilibrée entre tous les sujets d'intérêt, cet algorithme est clairement approprié au premier cas, quand le veilleur choisit des concepts précis pour formuler sa requête. Nos premiers algorithmes précédents qui reposaient sur des branches sont utiles pour le deuxième cas, puisqu'ils aident à découvrir les documents plus spécialisés.

8.4 Conclusion

Dans ce chapitre, nous avons évalué la composante la plus importante dans le système de veille OntoWatch : le module de recherche et génération des annotations sur le Web, en comparant les résultats de la veille avec et sans l'aide du système. L'évaluation a montré l'apport de l'ontologie dans la recherche d'information. Nous avons analysé les résultats obtenus pour expliquer les différents cas. En fait, le niveau d'exploitation de l'ontologie dans les algorithmes dépend de la structure arborescente de l'ontologie et du choix des concepts par l'utilisateur.

Pour une expérimentation plus complète à plus grande échelle, nous pouvons envisager de configurer l'algorithme pour travailler avec plus de documents dans les résultats de Google. Cependant, cela requerra évidemment plus de temps et de personnes dans la tâche d'analyse des documents trouvés par Google. Il faut déterminer si un document trouvé dans les Max premiers documents par la recherche guidée par l'ontologie, est situé aussi dans les résultats de la recherche normale mais très loin selon l'ordre de rangement des résultats de Google. L'analyse du contenu de ce document permet de savoir si Google aurait pu trouver

un nouveau document sans besoin de l'apparition des concepts descendants dans la requête.

Par contre, il est nécessaire d'étendre le nombre de documents traités par l'algorithme pour que le ratio $Nouveauté_{M/A}$ reste inférieur à un seuil et donc que la recherche automatique manque le moins possible des documents pertinents trouvés par la recherche normale.

En discutant avec les documentalistes du CSTB, il s'est avéré qu'il existe aussi différents niveaux de pertinence pour un document. Le contenu de plusieurs documents dans les résultats de recherche n'est pas très intéressant par rapport au sujet de veille, mais permet de naviguer vers d'autres documents plus pertinents. Par exemple, on cherche des rapports de recherche sur le sujet de «Web Sémantique » et le document trouvé n'est qu'un page web d'un auteur, mais cette page web contient des liens vers des rapports très pertinents sur ce sujet. Il faut donc ajouter des paramètres supplémentaires pour avoir des évaluations plus complètes.

Conclusion et perspectives

De plus en plus d'informations sont disponibles sur le Web. Au cours des dernières années, la capacité d'exploitation efficace de cette précieuse matière est devenue plus importante dans les activités de veille technologique et scientifique dans les entreprises et les organisation. Proposer une méthodologie et un support technique pour développer un système d'information aidant les veilleurs lors de leur veille sur le Web, tel est le principal objectif de cette thèse.

Confirmation des hypothèses initiales

On se trouve encore face aux problèmes de recherche d'informations plus précises et de résultats plus compréhensibles pour les veilleurs. En réponse à cette problématique, nous avons adopté les hypothèses de travail suivantes:

- Les technologies du Web Sémantique peuvent aider à construire un système de veille facilitant les tâches de recherche et de traitement d'informations par le veilleur.
- L'utilisation de l'ontologie peut être utile dans plusieurs étapes du processus de veille, et promet d'apporter de meilleurs résultats par rapport aux approches courantes qui reposent sur l'utilisation des outils de recherche d'information.

Le système OntoWatch proposé nous permet de confirmer nos hypothèses de départ. Ce système est inspiré des technologies du Web sémantique en gérant l'information et les résultats de veille par des annotations sémantiques en RDF. Ces annotations utilisent les termes définis dans l'ontologie O'Watch pour décrire les documents, les sources d'information, les profils d'utilisateur. Grâce au moteur de

recherche Corese, OntoWatch permet aux veilleurs d'effectuer des recherches sémantiques disposant de capacités d'inférence sur ces annotations. La prise en compte des distances entre concepts permet aux veilleurs d'obtenir non seulement des réponses exactes mais aussi des réponses approchées.

Sans compter que l'ontologie est un facteur indispensable dans la famille Web sémantique, nos études ont montré que l'utilisation de l'ontologie pour guider les tâches de veille constitue un réel apport, notamment grâce à la possibilité d'étendre la recherche normale sur le Web via les moteurs de recherche tel que Google, Yahoo, en bénéficiant de la structure hiérarchique de l'ontologie et les relations entre les concepts. Nous pouvons constater que l'ontologie est vraiment le cœur du système OntoWatch.

Nos Contributions

Du point de vue scientifique, plusieurs points peuvent être considérés comme innovants. Nos apports se situent dans plusieurs domaines de recherche:

- *Application de l'ontologie et des technologies du Web Sémantique à la veille technologique et scientifique*. Les recherches sur la construction des systèmes de veille reposent souvent sur les techniques de TALN (Traitement Automatique de la Langage Naturelle) ou des systèmes d'agents autonomes. Mais il existe encore peu de recherches sur la veille s'appuyant sur le Web Sémantique. Cette thèse met en valeur le rôle de l'ontologie dans le processus de veille du CSTB que nous avons modélisé en reposant sur le modèle générique de Lesca. Une ontologie dédiée à la veille technologique a été construite, à la fois grâce à notre travail de traduction des termes appartenant à des thésaurus, ce qui a conduit à des expériences intéressantes, et grâce à la réutilisation de certaines parties d'une ontologie existante.

- *Proposition des algorithmes pour la recherche de documents sur le Web et pour la génération automatique des annotations sémantiques.* Ces algorithmes utilisent l'ontologie (pour enrichir la recherche initiale du

veilleur sur le Web). Plus précisément, ils exploitent (a) la structure de l'ontologie, (b) les relations de subsomption entre les concepts dans l'ontologie, et (c) les synonymes attachés à chaque concept. Au niveau de la recherche d'information, d'une part OntoWatch est capable de trouver des documents pertinents non retrouvés par la recherche simple via Google, d'autre part la précision de la recherche est ainsi augmentée. Au niveau de la génération des annotations, nous avons proposé une méthode automatique pour alimenter les bases d'annotations sémantiques RDF du système, en plus des méthodes manuelles et semi-automatiques. Notre méthode diffère des méthodes s'appuyant sur l'apprentissage presenté dans Amilcare et MnM qui exige une phase d'entrainement sur les exemples de documents annotés. L'intégration de la génération des annotations RDF avec les résultats de recherche en utilisant l'ontologie, permet de fournir un grand nombre d'annotations, qui, malgré leur faible niveau de complexité, aident à réduire la charge lourde des veilleurs.

- *Proposition d'une architecture multi-agents coopérant pour implémenter OntoWatch.* Les techniques de la recherche sémantique, les algorithmes proposés sont encapsulés dans des agents spécifiques appartenant à différentes sociétés d'agents. Cette architecture est flexible et ouverte pour accepter sans effort de nouveaux rôles d'agents pour les fonctionnalités additionnelles. Le choix d'une architecture multi-agents pour développer OntoWatch est applicable et efficace pour prendre en compte la distribution des tâches des utilisateurs et l'hétérogénéité des sources d'information.

Limites et Perspectives

L'approche sémantique que nous avons proposée s'appuie sur des annotations sémantiques et des ontologies. Naturellement, la performance et l'efficacité du système dépendent beaucoup de la richesse de l'ontologie et de la qualité des

annotations sémantiques. Comme nous l'avons remarqué dans les chapitres précédents, une condition préalable pour assurer un bon résultat de la recherche sur le Web effectuée par les algorithmes utilisant l'ontologie est la richesse à la fois des concepts, et des relations entre concepts. De plus, dans le domaine de la veille, ce qui est important pour le veilleur c'est la nouveauté : donc l'ontologie, en tant que ressource conceptuelle fournissant des primitives sémantiques, doit évoluer avec le temps. La prise en compte des méthodes pour gérer l'évolution de l'ontologie n'a pas été approfondie dans notre travail.

Une autre limite de ce travail reste la richesse au niveau du contenu des annotations sémantiques. L'automatisation permet aux algorithmes de réduire le temps et les efforts humains, cependant ils doivent satisfaire des critères prédéfinis et à un certain niveau de généralité, la richesse des annotations n'est pas comparable avec celle des annotations obtenues par des méthodes d'annotation manuelle ou semi-automatique. Concrètement, dans l'état actuel, ces algorithmes ne génèrent que des annotations utilisant des concepts dans l'ontologie et certaines relations simples entre eux.

Perspectives à court et moyen terme

Le travail accompli à ce jour pourra être complété sur certains points. Les perspectives liées à la réalisation informatique sont:

- Implémentation de l'architecture multi-agents déjà conçue, en exploitant la plate-forme de développement des systèmes multi-agents JADE. Cette tâche consiste (1) à déterminer des comportements de chaque nouveau agent à partir de son rôle et ses interactions, (2) puis à implémenter ces comportements dans un type d'agent (classe d'agent) et (3) à intégrer dans OntoWatch des types d'agent existant dans CoMMA.
- Réalisation d'une évaluation à plus grande échelle sur la performance des algorithmes, concernant une analyse par les veilleurs du contenu d'un nombre plus élevé des documents obtenus dans les résultats de Google (ce qui nécessitera donc plus de temps de la part des veilleurs impliqués dans

cette évaluation). Une telle évaluation permettrait d'obtenir des résultats quantitatifs d'expérimentation plus précis sur la recherche normale et la recherche d'information utilisant l'ontologie.

Par rapport à une amélioration à moyen terme, il serait intéressant d'ajouter à la tâche de génération des annotations sémantiques les nouvelles fonctionnalités suivantes :

- Insérer à une annotation d'un document des informations obtenues par l'inférence sur les annotations sur les sources d'information. Cette tâche concerne la définition d'un ensemble de règles d'inférences concernant les sources existantes gérées par le système.
- Raffiner le module d'extraction de l'information en prenant en compte de petits indices de reconnaissance des informations utiles pour les annotations. Ces indices proviennent des expériences du veilleur.
- Classer des résultats d'annotation selon certains critères comme le thème concerné, la source dont sont issus les documents, pour aider les veilleurs dans la tâche d'analyse et de validation des annotations.

Perspectives à long terme

Rappelons notre intérêt pour le problème de la qualité des annotations générées automatiquement. Pour pouvoir mieux exploiter les informations cachées dans des données textuelles sur le Web, nous ne pouvons pas oublier les travaux linguistiques. L'intégration des outils de Traitement Automatique de la langue Naturelle TALN dans la tâche d'extraction d'information promet la possibilité d'annoter des documents avec des relations entre les concepts. Ce travail a été mené dans le cadre d'une autre thèse d'ACACIA [Khelif, 2006]. Le problème est d'intégrer les outils linguistiques à OntoWatch. L'architecture multi-agents souple d'OntoWatch permet de faciliter l'ajout de nouvelles fonctionnalités comme le traitement linguistique. Cependant il faut offrir la capacité d'intégrer ces outils TALN, ou bien de les encapsuler dans des agents autonomes.

Enfin, il est essentiel de développer des agents attachés à certaines sources considérées par les veilleurs comme importantes. Lors de la publication d'un nouveau document sur la source observée, ces agents pourront prendre l'initiative de différentes actions comme informer le veilleur, ou générer automatiquement une annotation.

Bibliographie

[Afnor, 1998] AFNOR, 1998, *Prestation de veille et prestation de mise en place d'un système de veille.*

[Amann and Fundulaki, 1999] Amann B., Fundulaki I., 1999. *Integrating Ontologies and Thesauri to Build RDF Schemas*. In ECDL-99: Research and Advanced Technologies for Digital Libraries, Lecture Notes in Computer Science, pages 234--253, Paris, France. Springer-Verlag.

[Arpírez et al., 2003] Arpírez J. C., Corcho O., Fernandez-López M., Góméz-Pérez A., 2003. *WebODE in a nushell*. AI Magazine 24(3) : 37-47, 2003.

[Aussenac-Gilles et al., 2000] Aussenac-Gilles N., Biebow B., Szulman S, 2000. *Revisiting ontology design: a method based on corpus analysis*, in Proceedings of the conference EKAW'2000, Springer LNCS 1937, pages 172-188, 2000.

[Bachimont, 2001] Bachimont B., *"Modélisation linguistique et modélisation logique des ontologies: l'apport de l'ontologie formelle."* In Proceedings of IC 2001, pp 349-368 Plate-forme AFIA, Grenoble 25-28 juin 2001

[Bauer and Leake, 2002] Bauer, T. and Leake, D., 2002 Calvin: *A Multi-Agent Personal Information Retrieval System*, Agent Oriented Information Systems 2002: Proceedings of the Fourth International Bi-Conference Workshop AOIS-2002

[Baumard, 1991] Baumard P., 1991, *Stratégie et surveillance des environnements concurrentiels*. Massons, Paris

[Bellifemine et al, 2001] Bellifemine F., Poggi A., Rimassa G., *Developing multi agent systems with a FIPA-compliant agent framework*. Software Practice & Experience, (2001) 31:103-128

[Blazquez et al., 1998] Blazquez M., Fernandez M., Garcia-Pinar J. M., Gomez-Perez A., *Building Ontologies at the Knowledge Level using the Ontology Design Environment*, in Proceedings of the Banff Workshop on Knowledge Acquisition for Knowledge-based Systems, 1998.

[Boissier et Demazeau, 1996] Boissier, O. and Demazeau, Y. (1996). *Asic: An architecture for social and individual control and its application to computer vision*. In John W. Perram and Jean-Pierre Müller, editor, Distributed Software Agents and Applications, 6th European Workshop on Modelling Autonomous Agents - MAAMAW '94, volume 1069, pages 1–18, Denmark. Springer.

[Bouquet et al., 2003] Bouquet P., Giunchiglia F., van Harmelen F., Serafini L., and Stuckenschmidt H., *C-OWL: Contextualizing Ontologies*. International Semantic Web Conference 2003: 164-179.

[Cao et al., 2006] Cao T-D., Dieng-Kuntz R., Fiès B., Bourdeau., *Vers un système d'aide à la veille technologique guidé par une ontologie* RFIA'2006, Tours, January 25-27, 2006.

[Cao et al., 2004] Cao T-D., Dieng-Kuntz R., Fiès B., *An Ontology-Guided Annotation System for Technology Monitoring*, IADIS International WWW/Internet 2004 Conference, Madrid, Spain, 6-9 October 2004.

[Cao et al, 2003a] Cao T-D., Gandon F., *Integrating external sources in a corporate semantic web managed by a multi-agent system*. Proc. of AMKM 2003, AAAI Spring Symposium on Agent-Mediated Knowledge Management March 24-26, 2003, Stanford University

[Cao et al, 2003b] Cao T-D., Gandon F., Dieng-Kuntz R., - *Intégration de sources extérieures dans un Web sémantique d'entreprise géré par un système multi-agents*, Actes des 14èmes journées francophones d'Ingénierie des Connaissances (IC'2003), plate-forme AFIA'2003, Laval, 1-3 juillet 2003, PUG.

[Ciravegna, 2001] Ciravegna F., (2001), *Adaptive Information Extraction from Text by Rule Induction and Generalisation*, in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle.

[Collier, 2002] Collier N., Takeuchi K., (2002), *PIA-Core: Semantic Annotation through Example-based Learning*, in proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, 29th – 31st May, pp. 1611-1614

[CoMMA, 2001] *Contribution au livrable CoMMA*, Unité documentaire du CSTB 22 mai 2001.

[eBib, 2000] Rapport final du projet e-BIB, 20-06-2000.

[Cote et Troudi, 1998] Côté M., et Troudi N., *"NetSA, Une Architecture Multiagent pour la Recherche sur Internet"*, Expertise Informatique, vol. 3(3), 1998

[Cote et al., 2001] Côté M., Chaib-draa B., Troudi N. *"NetSA : une architecture multiagent réutilisable pour les environnements riches en informations"*, Information, Interaction, Intelligence, Cépaduès, Toulouse, 1(2), pp. 39-78, 2001

[Croft et Cook, 1995] Croft W.B., Cook R, 1995. *Providing government information on the internet: Experiences with Thomas*. In Digital Libraries Conference, pages 19–24.

[DAML, 2000] DAML. *Darpa Agent Markup Language*. http://www.daml.org/about.html, 2000.

[Dean et al., 2003] Dean M., Schreiber Guus., van Harmelen F., Hendler J., Horrocks I., McGuinness D.L., PatelSchneider P. F., and Stein L.A., *OWL Web Ontology Language Reference*. 2003. http://www.w3.org/TR/owl-ref/

[Demazeau and Costa, 1996] Demazeau, Y. and Costa, A. R. (1996). *Populations and organisations in open multi-agent systems*. In 1st Symposium on Parallel and Distributed AI, Hyderabad, India

[Dieng et al., 2003] Dieng-Kuntz R., Corby O., Gandon F., Giboin A., Golebiowska J., Matta N., Ribière M., *Knowledgement management - Methodes Et Outils Pour La Gestion Des Connaissances* (3ème Edition), Dunod Edition - INFORMATIQUES Série Systèmes d'information

[Dou et Jakobiak, 1995] Dou H., Jakobiak F., *« De l'information documentaire à la veille technologique pour l'entreprise : enjeux, aspects généraux et définitions »*, in Veille technologique et compétitivité, Dunod, 1995.

[Drogoul, 1993] Drogoul, A. (1993). *De la Simulation Multi-agents à*

la Résolution Collective de Problèmes, Thèse de doctorat, Université Paris VI.

[Farquhar et al., 2000] Farquhar A., Fikes R., Rice J. (2000) *Ontolingua server : a tool for collaborative ontology construction*, in International journal of Human-Computer studies (46), pages 707-727, 2000.

[Fensel et al. 2000] Fensel D., van Harmelen F., Horrocks I., McGuinness D.L., Patel-Schneider P.F, 2001 *OIL: An Ontology Infrastructure for the Semantic Web*. IEEE Intelligent Systems 16(2): 38-45 (2001)

[Ferber, 1995] Ferber, J. (1995). *Les Systèmes Multi-Agents : vers une intelligence collective*. InterEditions.

[Ferber et Gutknecht, 1995] Ferber J., Gutknecht O. (1997) *Aalaadin : a meta-model for the analysis and design of organizations in multi-agent systems*, Rapport de Recherche LIRMM 97189

[Fernandez et al., 1997] Fernandez M., Gomez-Perez A., Juristo N., *METHONTOLOGY : from ontological art towards ontological engineering*, in Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97), AAAI Press , 1997.

[Gandon, 2002] Gandon F., (2002). *Distributed Artificial Intelligence and Knowledge Management: Ontologies and Multi-Agent Systems for Corporate Semantic Webs*, PhD Thesis, UNSA, 2002.

[GATE] GATE – *General Architecture for Text Engineering* http://gate.ac.uk

[Goujon, 2000] Goujon B., 2000, *Utilisation de l'exploration contextuelle pour l'aide à la veille technologique, Réalisation du système VIGITEXT*. Thèse de doctorat, Université Paris IV-Sorbonne.

[Gruber, 1993] Gruber TR., 1993, *Toward principles for the design of ontology's used for knowledge sharing*, Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, 1993.

[Gruniger et Fox, 1995] Gruninger M., Fox M. S., 1995, *Methodology for the design and evaluation of ontologies*, in Proceedings of the Workshop on Basic Ontological Issues on Knowledge Sharing, IJCAI'95.

[Guarino et Giaretta, 1995] Guarino N., Giaretta P., 1995, *Ontologies and Knowledge Bases Towards a Terminological Clarification*. In N. Mars (ed.) Towards Very Large Knowledge Bases : Knowledge Building and

Knowledge Sharing. IOS Press, Amterdam: 25-32.

[Guarino et Welty, 2000] Guarino N. & Welty C., *A Formal Ontology of Properties*, in Dieng R. & Corby O., eds., Knowledge Engineering and Knowledge Management : Methods, Models and Tools. International Conference EKAW'2000, Springer-Verlag, pages 97-112, 2000.

[Kassel, 2002] Kassel G., *OntoSpec : une méthode de spécification semi-informelle d'ontologies*, in Actes des journées francophones d'Ingénierie des Connaissances (IC'2002), pages 75-87, 2002.

[Handschuh, 2003] Handschuh S., Staab S., (2003) *CREAM Creating Metadata for the Semantic Web*. The International Journal of Computer and Telecommunications Networking.

[Hunt et Zartarian, 1990] Hunt C., Zartarian V., 1990, *Le renseignement stratégique au service de votre entreprise*. Edition First, Paris

[ISO 2788:1986] ISO (1986) ISO 2788:1986 *Documentation - Guidelines for the establishment and development of monolingual thesauri*. 2nd ed. (32 p.)

[Jakobiak 1995b] Jakobiak F., 1995 *Évaluation de la veille technologique*, in actes du colloque VSST'95, p. 253-271.

[Khelif, 2006] Khelif K., 2006 *Web sémantique et mémoire d'expériences pour l'analyse de transcriptome*, Thèse de doctorat, Université de Nice-Sophia Antipolis.

[Klusch, 2001] Klusch, M. 2001 *Information Agent Technology for the Internet: A Survey*. Journal on Data and Knowledge Engineering, Special Issue on Intelligent Information Integration, D. Fensel (Ed.), Vol. 36(3), Elsevier Science.

[Klusch, 1999] Klusch, M. 1999 *Intelligent Information Agents: Agent-Based Information Discovery and Management on the Internet.* Berlin: Springer-Verlag.

[Kushmerick, 2002] Kushmerick N., Thomas B., (2002), *Adaptive information extraction : core technologies for information agents*. In Intelligent Information Agents R&D in Europe: An AgentLink perspective. Springer.

[Laine, 1991] Laine F., 1991 *La veille technologique - De l'amateurisme au professionnalisme*. Eyrolles, 138 p.

[Lassila et Swick, 1999] Lassila O., Swick R., (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. World Wide Web Consortium. available at http://www.w3.org/TR/REC-rdf-syntax/.

[Laublet et al., 2002] Laublet P., Reynaud C., Charlet J. *Sur quelques aspects du Web sémantique*. Journée scientifique Web Sémantique Paris.

[Lesca et Blanco, 2002] Lesca H., Blanco S., 2002, *Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faible*, In 6ème Congrès international francophone sur la PME, HEC, Montréal.

[Lesca, 2002] Lesca H., 2002 *Veille Stratégique - Concepts et méthode de mise en place dans l'entreprise*. Third International Conference on Language Resource and Evaluation. Spain

[Lesca et Elisabeth, 1999] Lesca H., Lesca E., 1999, *Gestion de l'information, Qualité de l'information et performances de l'entreprise*, Editions du management, Paris.

[Martinet et Ribault, 1989] Martinet B., Ribault J.M, 1989 *La veille technologique concurrentielle et commerciale*. Les éditions d'organisation, p 389.

[Matthews et al., 2002] Matthews, B.M., Miller, K. and Wilson, M.D. (2002) *"A Thesaurus Interchange Format in RDF"*. Submitted to the Semantic Web Conference 2002.

[Noy et al., 2001] Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., Musen M. A. (2001). *Creating Semantic Web Contents with Protégé-2000*. IEEE Intelligent Systems, 16(2) :60-71.

[Occello et Demazeau, 1998] Occello, M. and Demazeau, Y. (1998). *Modelling decision making systems using agents for cooperation in a real time constraints*. In 3rd IFAC Symposium on Intelligent Autonomous Vehicles, volume 1, pages 51–56, Madrid, Spain.

[Popov, 2003] Popov B., Kiryakov A., (2003) *KIM – Semantic Annotation Platform*, 2nd International Semantic Web Conference (ISWC 2003)

[Qui et Frei, 1993] Qui, Y., Frei, H. (1993) *Concept based query expansion*, In: Proceedings of ACM SIGIR 1993 (16th Annual International ACM SIGIR Conference on Research and Development in Information

Retrieval), p. 160--169

[Rao et Georgeff, 1995] Rao, A. and Georgeff, M. (1995*). BDI agents : from theory to practice*. In conference of 1st International Conference on Multi-Agent Systems ICMAS, pages 312–319. AAAI Press.

[Rodriguez, 1994] Rodriguez, M. (1994). *Modélisation d'un agent autonome: Approche constructiviste de l'architecture de contrôle et de la représentation de connaissances*, Thèse de doctorat, Université de Neufchâtel.

[Rouach 1996] Rouach D., 1996, *La veille technologique et l'intelligence économique*, PUF.

[Samier et Sandoval, 1998] Samier H., Sandoval V., 1998, *La Recherche intelligente sur l'Internet - outils et méthodes*, Éditions Hermès, Paris.

[Singh et Huhns, 1999] Singh M.P., Huhns M. N., *Social Abstraction for Information Agents*, In Intelligent Information Agent: Agent-Based Information Discovery and Management on the Internet p37-52 Matthias Klusch Springer 1999

[Sure et al, 2002] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D., 2002, *OntoEdit: Collaborative ontology development for the Semantic Web*. Proceedings of the International Semantic Web Conference (ISWC).

[Uschold et King , 1995] Uschold M., King M., 1995, *Towards a methodology for building ontologies*, in Proceedings of IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing.

[Vargas, 2002] Maria Vargas, Enrico Motta et al (2002) *MnM Ontology driven tool for semantic markup*. In Proceedings of the workshop Semantic Authoring Annotation & Knowledge Markup.

[Weinstein et al., 1999] P.C. Weinstein, W.P. Birmingham, E.H. Durfee. *Agent-Based Digital Libraries: Decentralization and Coordination*. IEEE Communication Magazine, pp. 110-115, 1999

[Weiss, 1999] Weiss, G. (1999). *Multiagent systems and distributed artificial intelligence*. In Weiss, G., editor, Multiagent systems: A modern approach to Distributed Artificial Intelligence. MIT Press.

[Wooldridge, 1999] Wooldridge, M. (1999). Intelligent agents. InWeiss, G., editor, *Multiagent systems : A modern approach*

to Distributed Artificial Intelligence. MIT Press.

[Voorhees, 1994] Voorhees, E. M. (1994). *Query expansion using lexical-semantic relations*. Research and Development on Information Retrieval - ACM-SIGIR, Dublin, 61-70