UNIVERSITE KASDI MERBAH OUARGLA

FACULTE DES SCIENCES ET SCIENCES DE L'INGENIEUR

DEPARTEMENT DE MATHEMATIQUE ET D'INFORMATIQUE

N° d'ordre :

N° de série :

Mémoire

Présenté pour l'obtention du diplôme de

MAGISTER

Spécialité : Informatique

Option : Informatique et Communication Electronique

Par : GASMI Mounira

Thème

Utilisation des ontologies pour l'indexation automatique des sites Web en Arabe

Soutenu publiquement le : 27 mai 2009

Devant le jury composé de :

Mme. Fatima-Zohra LAALLAM	**MC. (Université de Ouargla)**	**Présidente**
Mr. Chabane KHENTOUT	**MC. (Université de Sétif)**	**Examinateur**
Mr. Samir ZIDAT	**MC. (Université de Batna)**	**Examinateur**
Mr. Mahieddine DJOUDI	**MC. (Université de Poitiers)**	**Rapporteur**
Mr. Lamri DOUIDI	**MC. (Université de Sétif)**	**Co-Rapporteur**

Remerciements

Je remercie tout d'abord mon Promoteur Monsieur Mahieddine DJOUDI, Maître de conférence à l'Université de Poitiers (France), et mon co-promoteur Monsieur Lamri DOUIDI Maître de conférence de l'Université de Sétif, pour tous les conseils et encouragements dont j'ai bénéficié tout au long de ce travail.

Mes respects et ma gratitude vont également à Mme Fatima-Zohra LAALAM Maître de conférence de l'Université Kasdi Merbah de Ouargla d'avoir accepter de présider le jury. J'adresse mes remerciements à Mr Samir ZIDAT Maître de conférence à l'Université de Batna, et Mr Chabane KHENTOUT Maître de conférence à l'Université de Sétif, qui m'ont fait l'honneur de juger ce travail

Je tiens à exprimer ma profonde gratitude à Mr Abd ElHakim HAROUZ pour tous les efforts faits pour la promotion de magister ICE 2005.

Par leurs conseils avisés, Mr Sliman BELLAOUAR et Mr Salim MEFLAH m'ont permis de préciser et d'améliorer mon travail. Je tiens à leur témoigner toute ma gratitude.

Je remercie également tous mes collègues de promotion du majister ICE de l'Université de Ouargla.

Je souhaite exprimer toute ma reconnaissance à Madame Hadda HAMMOUCHE (Mme GUERBATI) pour tout le soutien surtout moral durant le déroulement de ce mémoire.

Un remerciement spécial à mes parents et ainsi toute ma famille et ma belle famille pour leur soutien et compréhension tout au long de ce projet surtout ma sœur Wahiba.

Je ne saurais terminer ces remerciements sans un énorme merci à mon mari Azeddine MEHAOUCHI qui m'a supporté et m'a soutenue avec patience durant le déroulement du PG.

Un pardon à ma petite fillette Tasnim pour mon occupation totale pendant ces derniers jours.

A tous un sincère et chaleureux merci !

Résumé

La croissance du Web est entrain de faire une énorme masse d'information universelle. Le Monde Arabe, ces derniers temps, a contribué à cette explosion. Pour cette raison, il serait raisonnable de penser à des techniques efficaces qui permettraient à l'utilisateur arabe de trouver les documents pertinents qu'il cherche dans le Web.

Ce mémoire présente une approche d'indexation des sites web arabes, par l'utilisation des ontologies et les techniques de traitement automatique de la langue arabe pour la recherche d'information sur internet.

Tout d'abord une ontologie arabe (أنطولوجيا_جامعة) orientée terminologie de domaine est construite pour être utilisée dans le processus d'indexation. Nous avons utilisé un thésaurus linguistique (WordNet) couplé avec un dictionnaire bilingue (Tarjim de Ajeeb), dans un but de désambiguïsation des concepts d'une telle ontologie.

Notre outil d'indexation s'appuie sur les techniques issues du traitement automatique de la langue arabe (TALA) pour générer des termes bien formés à partir des pages web arabes. Les marqueurs HTML sont considérés. Ensuite les concepts associés aux termes bien formés sont générés par le biais d'un thésaurus. En fin l'index structuré est déduit par la mise en correspondance des concepts des pages web déterminés et les concepts de notre ontologie orientée terminologie.

Mots clés :

Indexation sémantique, Ontologie, OWL, Pages web arabes, Système de recherche d'information, traitement automatique de la langue arabe, Web sémantique.

ABSTRACT

The growth of the Web is making an enormous mass of a universal information. The Arab world, in recent times, has contributed to this explosion. For this reason, it would be reasonable to think about effective techniques that would enable the Arab user to find relevant documents that searches in the Web. This thesis presents an indexing approach of Arab websites, using ontologies and Arabic automatic natural Language processing for information retrieval on the Internet.

First an Arabic domain ontology (أنطولوجيا_جامعة) oriented terminology is built to be used in the indexing process. We used a linguistic thesaurus (WordNet) coupled with a bilingual dictionary (Tarjim of Ajeeb), in order to disambiguate concepts of such ontology.

Our indexing tool is based on techniques resulting from the Arabic automatic Language processing (AALP) to generate well-formed terms from Arabic web pages. HTML markers are considered. Then the concepts associated with well-formed terms are generated through a thesaurus. Finally, the structured index is deducted by the matching identified concepts of Web pages and the concepts of our arbic ontology oriented terminology.

Keywords:

Semantic indexation, Ontology, OWL, Arabic web Pages, Information retrieval systems, Arabic automatic Language processing, semantic Web.

ملخص

نمو شبكة الانترنت هو صنع كتلة هائلة من المعلومات عالميا. العالم العربي، في الآونة الأخيرة، قد أسهم في هذا الانفجار. ولهذا السبب، سيكون من المعقول التفكير في التقنيات الفعالة التي من شأنها أن تمكن المستخدم العربي من إيجاد الوثائق ذات الصلة التي يبحث عنها في الشبكة العالمية .

هذه الأطروحة تقدم نهج لفهرسة المواقع العربية باستخدام الأنتولوجيا و المعالجة الآلية للغة العربية من أجل استرجاع المعلومات على شبكة الإنترنت .

في البدء، أنطولوجيا عربية(أنطولوجيا_جامعة) موجهة المصطلحات تم بناؤها لاستخدامها في عملية الترميز، لقد استخدمنا المكنز اللغوي *(WordNet)* إلى جانب قاموس ثنائي اللغة (ترجم من عجيب) ، وذلك من أجل إزالة غموض مثل هذه الأنطولوجيا.

أداة الترميز تقوم على تقنيات المعالجة الآلية للغة العربية *(AALP)* لتوليد مصطلحات جيدة التركيب إنطلاقاَ من صفحات الويب العربية. وتؤخذ في عين الإعتبارعلامات *HTML*. بعد ذلك المفاهيم المرتبطة بالمصطلحات جيدة التركيب تتولد من خلال المكنز.

وأخيرا ، فإن الرمز المنظم يستخلص من خلال المطابقة بين المفاهيم المنتقاة من صفحات الويب والمفاهيم من الأنطوجيا موجهة المصطلحات.

الكلمات المفتاحية:

الترميزالدلالي، انطولوجيا، صفحات ويب عربية، نظم استرجاع المعلومات، المعالجة الآلية للغة العربية ، الويب الدلالي، ***OWL***.

Table des matières

Liste des figures

Liste des tables

Introduction générale

Introduction :

Les volumes astronomiques des données électroniques, la diversité et l'hétérogénéité des sources, durant les dernières décades, nécessitent une mise à niveau de la philosophie des traitements de ces données. Par exemple, un moteur de recherche populaire rapporte plus de huit (8) milliards de pages dans son index en juillet 2005 alors qu'elles étaient seulement 320 millions en 1997 et 3.3 milliards en septembre 2002.

Dans cette optique, plusieurs techniques ont été adoptées pour améliorer la performance et la pertinence des systèmes de recherche d'information (SRI). Parmi lesquelles nous citons : les techniques d'indexation basées sur les ressources sémantiques externes (ontologies, thesaurus…), et les techniques d'expansion de requêtes …etc.

En parallèle, vu que les données en langue arabe commencent à représenter une importante portion des données électroniques publiées (El-Hachani, 2005), des études spécialisées dans la recherche d'information en cette langue sont menées, plus particulièrement celles basées sur des traitements linguistiques : la racinisation, la pseudo-racinisation…

Notre travail consiste à contribuer dans l'amélioration des performances des SRIs Arabes sur le web en faisant recours aux techniques d'indexation basées sur les ontologies.

Motivation :

En plus de la progression rapide du web (contenu et internaute), nous assistons à une croissance exponentielle des sites web arabes, taux de croissance des internautes arabes entre 2000 et 2007 de 931,8%. Pour l'Algérie, avec une population de 33.506.567, le nombre des internautes en décembre 2000 était de 50.000, alors qu'en mars 2007 le nombre est devenu 1.920.000, soit un taux de croissance de 3.740,0%.

Alors La recherche d'information en langue Arabe est devenue de plus en plus importante. Néanmoins, peu de moteurs de recherche spécialisés en cette langue existent, d'où la nécessité de mener des recherches dans ce contexte.

A cette fin, nous nous sommes intéressés, dans le contexte de notre travail, à plusieurs disciplines : le traitement automatique des langues naturelles (en tenant compte des spécificités de la langue arabe), l'indexation, et l'utilisation des ontologies comme ressource externe lors du processus d'indexation.

Contribution :

Ce travail s'inscrit dans le cadre de la recherche d'information en langue arabe par l'utilisation les technologies du Web sémantique. Nous présentons dans ce qui suit les principales contributions de ce mémoire :

1. *Construction d'une ontologie de domaine* أنطولوجيا_جامعة: nous construisons une ontologie de domaine d'universités arabes. Pour des finalités d'indexation pour fournir une base de connaissances réutilisables à n'importe quelle université arabe. Nous suivons comme méthodologie nos inspirations du guide de Noy et McGuinness. Aprés, cette ontologie va être traité pour avoir une ontologie orienté terminologie, par le biais d'un processus de désambiguïsation de ses étiquettes (en utilisant un thésaurus linguistique).
2. *Présentation de l'outil d'indexation des pages web arabe* : elle est basée sur l'indexation du contenu des pages web par une ontologie arabe orientée

terminologie de domaine, et les techniques du traitement automatique de la langue arabe. Elle se compose de trois modules :

- Module d'extraction des termes arabes.
- Module de détermination des concepts.
- Module de construction d'index.

Conçus pour une finalité d'avoir un index structuré représentatif des pages web.

Organisation du mémoire :

Le mémoire est organisé selon trois chapitres :

Les deux premiers chapitres présentent l'état de l'art relatif aux domaines en lien avec nos travaux et le troisième décrit nos contributions :

Chapitre1 :

Nous dressons, un rapide état de l'art sur le web sémantique, son intérêt dans la représentation des connaissances, ses principales technologies et ses applications. Après, nous faisons un tour d'horizon autour des ontologies, en s'appuyant sur leurs types, leurs composants, leur construction et leur alignement et fusion, et enfin (on a terminé par) nous présentons son utilité dans la recherche d'information.

Chapitre 2 : est décomposé en 3 sections :

Dans la première section de ce chapitre les principales notions et concepts de la recherche d'information sont mentionnés. Après nous développons les principales étapes d'un processus de recherche d'information, nous faisons ensuite un panorama des outils de la recherche d'information. Par la suite, nous décrivons l'utilité de l'utilisation des ontologies dans le domaine de RI, enfin nous évoquons quelques-unes des grandes tendances qui peuvent résumer les principaux bouleversements de la recherche d'information.

Dans la deuxième section, nous dressons un état de l'art sur l'évolution de la place de la langue arabe dans le domaine de la recherche d'information sur le web, passant par la

définition de ses caractéristiques ainsi que le problème de son encodage et les solutions apportées. Après nous présentons les nouvelles recherches inhérentes à l'utilisation des ontologies avec la langue arabe.

Enfin dans la section 3, nous faisons un tour d'horizon sur l'indexation telle qu'elle se définit en recherche d'information. Commençant par la définition et les objectifs, passant par les types et les langages d'indexation, ensuite, nous abordons la partie d'indexation des documents sur le web, ou nous parlons d'analyse de document par la description de l'extraction des termes ou descripteurs, et la pondération où l'objectif est de trouver les termes qui caractérisent au mieux le contenu d'un document. A la fin de cette section, nous discutons les approches basées sur l'indexation des documents Web basées sur les ontologies.

Chapitre 3 :

Nous optons pour un choix de démarche pour l'indexation sémantique des pages web arabes. Après, les étapes de construction de l'ontologie de domaine pour des finalités d'indexation sont établies. Par la suite, décrivons l'architecture et la description de notre outil d'indexation des pages web arabes, cet outil adopte la démarche choisie antérieurement.

Finalement, nous concluons par un bilan de notre travail. Après, nous citons quelques perspectives.

Chapitre 1 : Web sémantique et ontologies.

1.1 Introduction :

Depuis sa création, il y a plusieurs années, par Tim Berners-Lee, le World Wide Web a révolutionné considérablement plusieurs domaines, notamment la société, l'économie et surtout la recherche et la manière dont celle-ci est conduite (Mestiri, 2007).

Le web est constitué par un ensemble de documents formatés dans le langage HTML, qui fournit particulièrement des liens hypertextes. Il sont exploités par des navigateurs ou rebots de recherche, tout ça est conçu pour être lu et compris par les humains, mais les dispositifs logiciels n'avaient aucune idée sur le sens.

L'idée du web sémantique, qui est une extension du web actuel, est de concevoir un web compréhensible par la machine pour une meilleure coopération homme machine.

Le web sémantique est fondé sur plusieurs technologies, comme RDF, OWL et XML qui ont été recommandés par le W 3C (World wide web consortium: c'est un groupement fondé en octobre 1994 pour élaborer des normes et émettre des directives visant la meilleure utilisation possible du Web dans la société.). Ces technologies ou langages constituent un socle pour définir des ontologies, entités servant à faciliter l'exploitation automatique ou semi-automatique du contenu par un ordinateur tout en gardant la signification du contenu pour les êtres humains (Ghafour, 2003).

1.2 Web sémantique :

« The Semantic Web is an extension of the current web in which information is given

well-defined meaning, better enabling computers and people to work in cooperation. » Tim Berners-Lee

1.2.1 Intérêt du web sémantique :

Tim BernersLee, a proclamé que le Web sémantique est la prochaine évolution du Web. C'est-à-dire que l'on va arriver à un Web intelligent où les informations sont stockées de façon compréhensible par les ordinateurs afin d'apporter à l'utilisateur ce qui cherche vraiment (Phan, 2005).

Parmi les points forts du web sémantique que :

- Il munit les langages du Web d'une sémantique formelle à l'aide d'une interprétation en terme d'un modèle. Elle permet une caractérisation précise des opérations applicables, par exemple de pouvoir affirmer la correction des algorithmes comme des algorithmes de recherche (Laublet et al., 2002).

- Les services seront mieux rendus sans engendrer de surcharge pour les utilisateurs (Baget et al., 2003).

- La recherche sur l'internet sera affine. Pour le faire, il va ajouter aux informations existantes une couche de métadonnées pour que les ordinateurs puissent l'exploiter (Phan, 2005). Cet ajout est du à l'insuffisance majeure du HTML : il ne sépare pas le contenu de la présentation, ce qui pose un problème d'interopérabilité (Mestiri, 2007).

- Il fait coopérer étroitement des acteurs d'origines très différentes: depuis les protocoles de communication jusqu'aux relations entre ordinateur et sens. On peut noter une double partition dans les développements du web sémantique:

  D'une part entre théoriciens et praticiens, d'autre part entre industriels et militants (Euzenat, 2003).

- Il définit des ontologies qui permettraient aux humains et aux machines de partager les connaissances du domaine et de collaborer ensemble. On les utilisant en général pour permettre aux machines de raisonner et d'interpréter les informations ainsi que d'améliorer la pertinence des recherches. Les agents auxquels les utilisateurs délègueront des tâches, devront communiquer entre eux et interpréter le contenu échangé de la même manière, c'est-à-dire en interprétant

les termes décrivant le contenu de la même manière. D'où l'intérêt de cette ontologie. « Avec la notion de Web sémantique, vous définissez un espace virtuel où les hyperliens pointeraient non plus sur des documents (textes ou images), mais sur des concepts » explique Pierre Lévy (Mestiri, 2007).

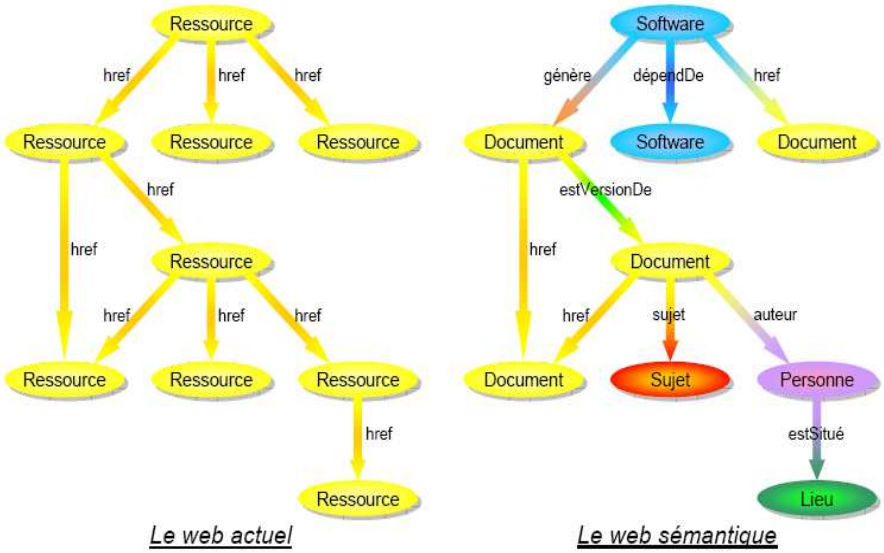

Figure 1. : Comparaison entre le web actuel et le web sémantique.

1.2.2 La représentation de la connaissance :

Pour que le Web sémantique fonctionne, les ordinateurs doivent avoir accès à des collections structurées d'informations et d'ensembles de règles d'inférence qu'ils peuvent utiliser pour parvenir à un raisonnement automatisé (Berners-Lee et al., 2001). Cette technologie souvent appelée représentation de la connaissance en est actuellement à un stade comparable à celui de l'hypertexte avant le Web.

Le web sémantique, contrairement aux systèmes traditionnels, accepte les paradoxes. Le défi du web sémantique est de fournir un langage qui exprime à la fois des données et des règles pour raisonner sur les données et pour que les règles de n'importe quel système de représentation de la connaissance puissent être exportées sur le Web (Berners-Lee et al., 2001).

C'est-à-dire on ajoute de la logique au Web pour lui donner la possibilité d'utiliser les règles pour faire des inférences, choisir des cours d'action et répondre aux questions de l'utilisateur (Phan, 2005).

Deux importantes technologies de développement du web sémantique XML et RDF.

RDF est utilisé pour coder l'information dans un document, et XML permet à chacun de créer ses propres balises.

1.2.3 Outils et technologies pour le web sémantique :

Les travaux visant la réalisation du Web sémantique se situent à des niveaux de complexité très différents. Les plus simples utilisent des jeux plus ou moins réduits de méta-données dans un contexte de recherche d'information ou pour adapter la présentation des informations aux utilisateurs. Dans ce cas, des langages de représentation simples sont suffisants. Dans les travaux plus complexes mettant en oeuvre des architectures sophistiquées, pour permettre par exemple l'exploitation de ressources hétérogènes, des langages plus expressifs et plus formels issus des travaux en représentation et en ingénierie des connaissances, sont nécessaires (Laublet et al., 2002).

Il semble clair que le web sémantique ne pourra voir le jour sans un minimum de standardisation. En février 2004, OWL et RDF sont devenus des recommandations du W3C. RDF est utilisé pour représenter l'information et pour échanger la connaissance sur le Web. OWL est utilisé pour publier et partager les ensembles de termes (appelés les ontologies) (Boutemedjet, 2004). Et aussi topic maps est une norme ISO (Baget et al., 2003).

Dans (Baget et al., 2003), les auteurs ont classé les langages du web sémantique selon trois sortes :

- Des langages d'assertions (RDF et cartes topiques) ;
- Des langages de définition d'ontologies (OWL, DAML+OIL) ;
- différents langages de description et de composition de services!(UDDI et autres).

Avant d'entamer ce classement, décrivons au début un langage de base, le langage XML, l'un des importantes technologies de développement du web sémantique.

1.2.3.1 Le langage XML :

Le développement de XML (eXtensible Markup Language) a commencé en 1996, et XML est une norme du W3C depuis février 1998 (Bos, 2002). XML est un ensemble de règles, de lignes directrices, de conventions (quel que soit le nom que vous voulez leur donner) pour la conception de formats texte permettant de structurer des données. XML facilite la réalisation de fichiers qui ne soient pas ambigus, et qui évitent les pièges courants, tels que la non extensibilité, l'absence de prise en charge de l'internationalisation/localisation et la dépendance par rapport à certaines plates-formes. XML est conforme à Unicode (Bos, 2002).

XML est un métalangage qui permet de définir d'autres langages. Les langages définis par XML sont des langages de présentation de documents (Phan, 2005). Il est donc naturellement utilisé pour encoder les langages du web sémantique (Une description de type de document, DTD, permet de décrire la grammaire des documents admissibles) (Baget et al., 2003).
Mais, ceci ne permet pas à une machine de manipuler sémantiquement un document, ainsi, une annotation sera attachée de la même manière à un paragraphe, un exposant dans une formule mathématique ou un polygone dans un dessin parce que ceux-ci sont encodés en XML.
Cette compatibilité entre les langages décrits en XML permet de construire les langages présentés ci-dessous et de les considérer comme des documents XML.

1.2.3.2 Langages d'assertions et d'annotations :

Les assertions affirment l'existence de relations entre des objets. Elles sont donc adaptées à l'expression des annotations que l'on veut associer aux ressources du web (Baget et al., 2003).

On va présenter principalement : le langage RDF et le formalisme cartes topiques.

1.2.3.2.1 RDF (Resource Description Framework) :

RDF est un langage formel (Baget et al., 2003) qui permet de voir le Web comme un ensemble de ressources reliées par les liens étiquetés « sémantiquement », il permet aussi d'exprimer de larges vocabulaires (Laublet et al., 2002). RDF sera utilisé pour annoter des documents écrits dans des langages non structurés, ou comme une interface pour des documents écrits dans des langages ayant une sémantique équivalente (des bases de données, par exemple) (Baget et al., 2003).

Un document RDF est un ensemble de triplets de la forme <*sujet, prédicat, objet*>.Les éléments de ces triplets peuvent être des URIs (Universal Resource Identifiers), des littéraux (Littéral : est un objet qui n'est pas une URI mais bien un contenu réel, une valeur). Ou des variables. Cet ensemble de triplets peut être représenté de façon naturelle par un graphe (plus précisément un multiplegraphe orienté étiqueté), où les éléments apparaissant comme sujet ou objet sont les sommets, et chaque triplet est représenté par un arc dont l'origine est son sujet et la destination est son objet. Ce document sera codé en machine par un document RDF/XML, mais qui est souvent représenté sous une forme graphique (Beckett, 2003 ; Phan, 2005).

Le sujet : Cela peut être n'importe quel objet référencé par une URI, qu'il concerne le web (Page HTML, document PDF, fichier multimédia…), ou non (Personne, Région, Etc.).

Le prédicat : Critère, caractéristique, attribut ou relation qui peut décrire la ressource (titre, couleur, taille, auteur, etc.). Une propriété ses valeurs permises et ses relations avec les autres propriétés.

L'objet : C'est la valeur qui sera affectée à la propriété de la ressource. Cette affectation peut être soumise à certaines restrictions (Mestiri, 2007).

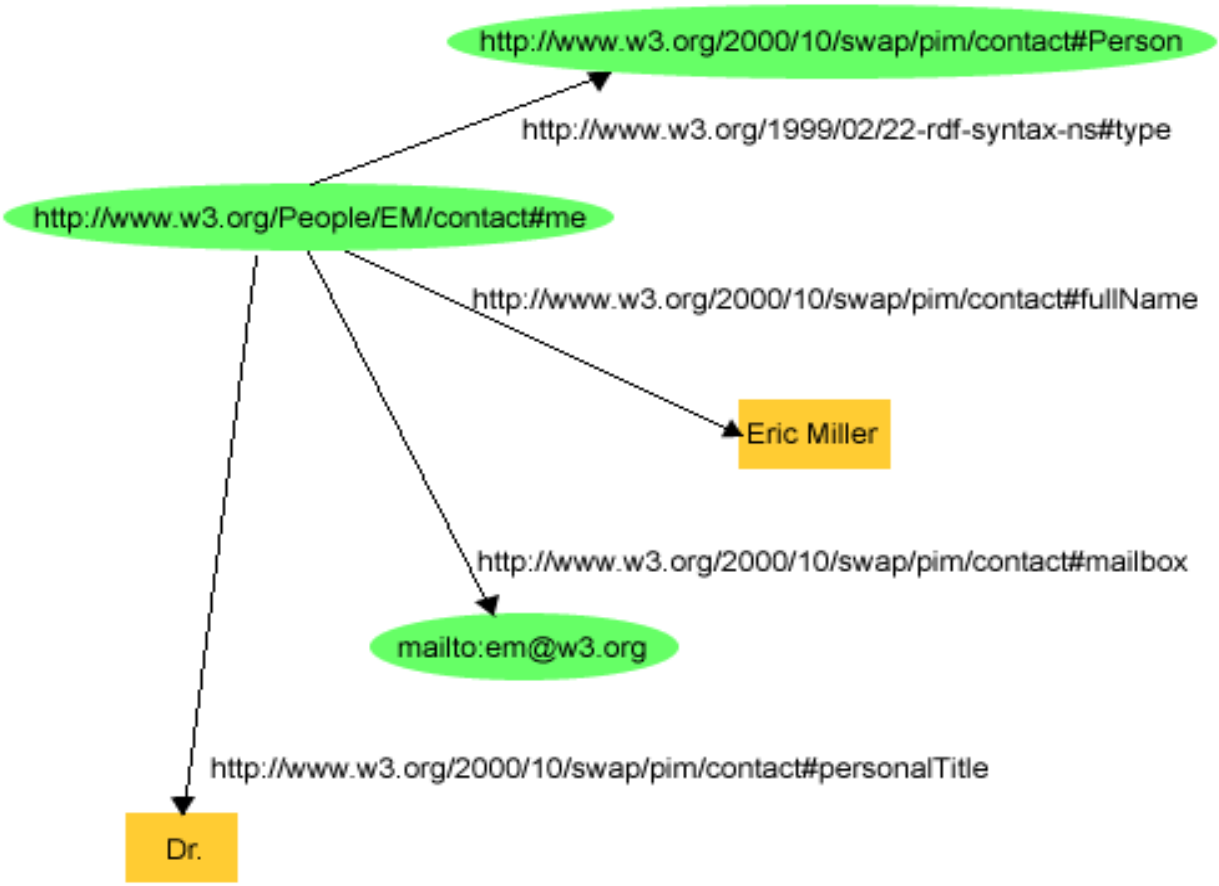

Figure 2. : Une illustration du RDF.

Un exemple est proposé en figure 2. La personne Eric Miller

- est identifiée par l'URI http://www.w3.org/People/EM/contact#me
- est de type (http://www.w3.org/1999/02/22-rdf-syntax-ns#type)

  Personne (http://www.w3.org/2000/10/swap/pim/contact#Person)
- a un nom complet

  (http://www.w3.org/2000/10/swap/pim/contact#fullName) Eric Miller
- a une adresse électronique

  (http://www.w3.org/2000/10/swap/pim/contact#mailbox) em@w3.org
- et a un titre

  (http://www.w3.org/2000/10/swap/pim/contact#personalTitle) Dr.

Les valeurs Eric Miller et Dr. sont encadrées d'un rectangle, car contrairement aux autres éléments de ce graphe RDF, ils ne sont pas des URI, mais bien des littéraux (McBride et Packard, 2004).

RDFS (pour RDF Schéma) a pour but d'étendre ce langage en décrivant plus précisément les ressources utilisées pour étiqueter les graphes. Pour cela, il fournit un mécanisme permettant de spécifier les classes dont les ressources seront des instances, comme les propriétés (Beckett, 2003).

RDFS s'écrit toujours à l'aide de triplets RDF, en définissant la sémantique de nouveaux mots clés comme notés dans (Baget et al., 2003):

<*ex:Vehicule rdf:type rdfs:Class*> la ressource *ex:Véhicule* a pour type *rdfs:Class*, est donc une classe;

<*sncf:TER8153 rdf:type ex:Vehicule*> la ressource *sncf:TER8153* est une instance de la classe *ex:Véhicule* que nous avons définie ;

<*sncf:Train rdfs:subClassOf ex:Vehicule*> la classe *sncf:Train* est une sous_classe de *ex:Véhicule*, toutes les instances de *sncf:Train* sont donc des instances de *ex:Véhicule;*

<*ex:localisation rdf:type rdfs:Property*> affirme que <*ex:localisation*> est une propriété (une ressource utilisable pour étiqueter les arcs);

<*ex:localisation rdfs:range ex:Ville*> affirme que toute ressource utilisée comme extrémité d'un arc étiqueté par *ex:localisation* sera une instance de la classe *ex:Ville.*

Mais avec RDFS, on a les problèmes montionnés dans (Phan, 2005):

- Il est assez faible pour décrire des ressources de façon plus détaillées. En plus, il ne sait pas quels sont les types des contraintes correspondants avec quel domaine. Par exemple : la propriété *hasChild* est une personne quand on l'applique aux domaines des Personne, mais elle est un petit éléphant sur les domaine des Éléphants.

- Il n'a aucune contrainte d'existence et de quantité. Par exemple: Si on a tous les instances de personne ont une mère, donc, elle est aussi une Personne ou nous, les personnes, avons seulement un parents, le père et la mère .etc.
- Il n'y a pas les propriétés transitives, inverses, ou symétriques. Par exemple: la propriété *isPartOf* est transitive, ou *hasPart* est inverse de celui. Et plus, il est assez complexe pour un raisonnement.

1.2.3.2.2 Cartes topiques :

Cartes topiques ou topic maps est une proposition concurrente à RDF(S) pour représenter les méta-données. Une première standardisation, fondée sur une DTD SGML, en a été donnée par (ISO). L'approche des Topic Maps repose sur les notions de topics qui peuvent être n'importe quel sujet ou entité, d'associations qui étiquettent des relations entre topics et d'occurrences qui sont des ressources, disposant d'une URI, qui peuvent être liées à des topics (Laublet et al., 2002).

Les "topics" que l'on peut comprendre comme des individus des langages de représentation de connaissances.

Les noms donnés aux topics: l'une des originalités des cartes topiques est la séparation des concepts et de leurs noms. Cela permet d'avoir plusieurs noms pour le même concept (et donc d'avoir des cartes topiques multilingues) et des noms partagés par plusieurs concepts (Baget et al., 2003).

Par exemple, le topic de vol est instancié par myFlight, il a pour nom «vol pour Boston» dont la portée est celle de mes discussions au déjeuner avec les collègues et "flight AF322" lors de discussions avec l'immigration américaine.

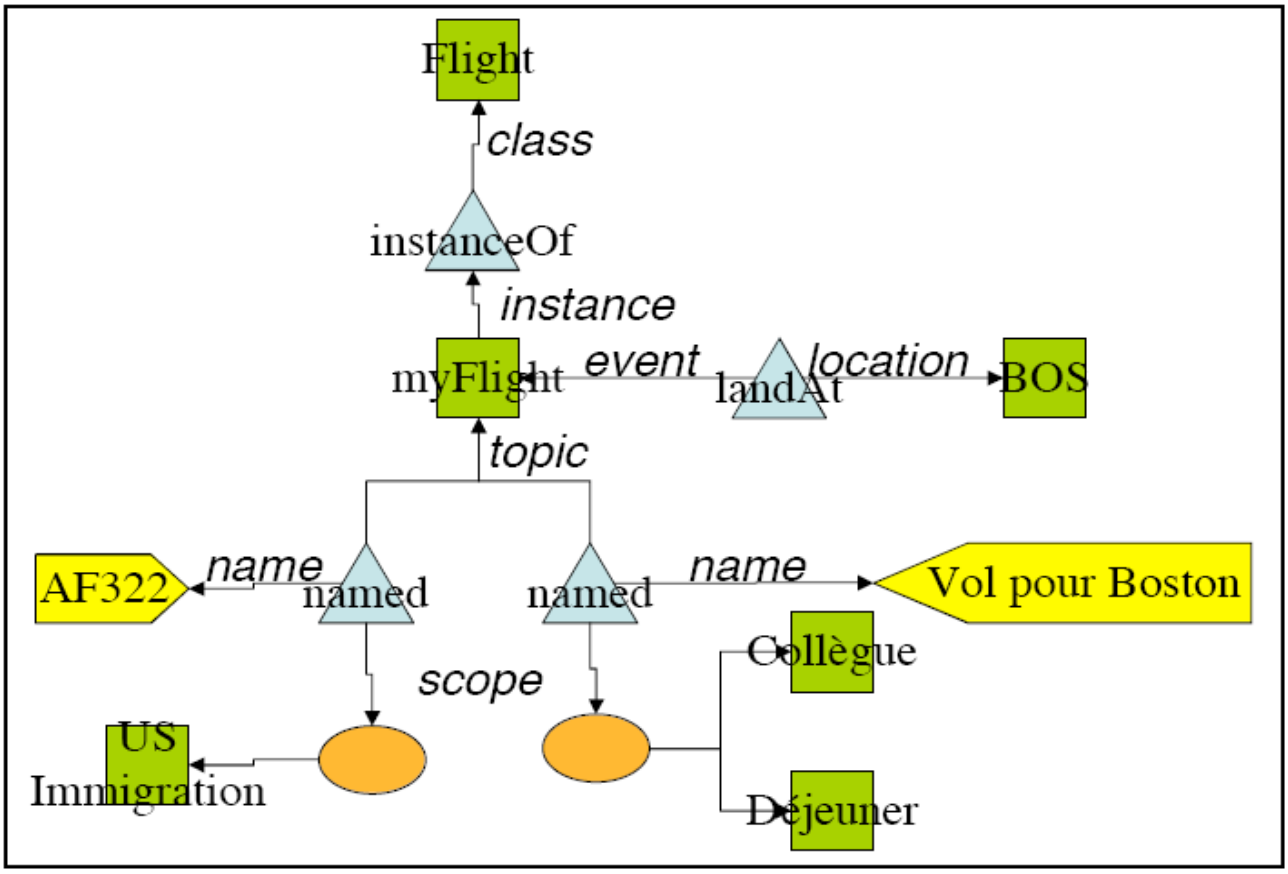

Figure 3. : Illustration de l'exemple pour Topic Maps.

1.2.3.3 Langages de définitions d'ontologies :

RDF, langage dédié à l'expression d'assertions sur les relations entre objets, s'est heurté à la nécessité de définir les propriétés des classes dont ces objets sont instances. Cependant, l'extension à RDFS ne fournit que des mécanismes très basiques pour spécifier ces classes. Le langage OWL, quant à lui, est dédié aux définitions de classes et de types de propriétés, et donc à la définition d'ontologies. Inspiré des logiques de descriptions (il est successeur de DAML+OIL) (Baget et al., 2003).

1.2.3.3.1 DAML+OIL :

Beaucoup de travaux ont été faits dans le domaine de la représentation des connaissances parmi lesquels on peut citer les plus importants : SHOE, OntoBroker, OIL (Fensel et al., 2000), et encore DAML + OIL (Connolly et al., 2001) qui a remplacé DAML – ONT (Ghafour, 2003).

DAML + OIL est un langage construit sur des normes précédentes du W3C telles que RDF et RDF Schéma, et étend ces langages avec des primitives de modélisation plus riches. DAML+OIL a été conçu à partir du langage d'ontologie DAML-ONT (DARPA Agent Modelling Language-Ontology, Octobre 2000) en vue de combiner plusieurs composants du langage OIL (Fensel et al., 2000). OIL « *Ontology Inference Language* » est une représentation basée sur le Web, et une couche d'inférence pour des ontologies. Il combine les primitives de modélisation des langages à base de cadres (frames) avec la sémantique formelle et le raisonnement fournis par la logique de description (Ghafour, 2003).

1.2.3.3.2 OWL (Ontology web Language) :

Le W3C a mis au point OWL (Web Ontology Language ou Langage d'Ontologie Web) [owl], conçu pour étendre RDF et préciser les ontologies. Il enrichit le modèle des RDF Schémas en définissant un vocabulaire riche pour la description d'ontologies Web structurées (et complexes) (Schuurman, 2005). Bien précis, OWL peut être utilisé pour représenter explicitement les sens des termes des vocabulaires et les relations entre ces termes. OWL vise également à rendre les ressources sur le Web aisément accessibles aux processus automatisés (McGuinness et Harmeln, 2004).

Le langage OWL est assez complexe, voilà pourquoi il se compose de trois sous langages qui proposent une expressivité croissante, chacun conçu pour des communautés de développeurs et des utilisateurs spécifiques : OWL Lite, OWL DL, OWL Full. Chacun est une extension par rapport à son prédécesseur plus simple (Schuurman, 2005).

- Le langage OWL Lite répond à des besoins de hiérarchie de classification et de fonctionnalités de contrainte simples de cardinalité 0 ou 1. Une cardinalité 0 ou 1 correspond à des relations fonctionnelles, par exemple, une personne a une adresse. Toutefois, cette personne peut avoir un ou plusieurs prénoms, OWL Lite ne suffit donc pas pour cette situation.

- Le langage OWL DL concerne les utilisateurs qui souhaitent une expressivité maximum couplée à la complétude du calcul (cela signifie que toutes les

inférences seront assurées d'être prises en compte) et la décidabilité (c'est-à-dire que tous les calculs seront terminés dans un intervalle de temps fini) du système de raisonnement. Ce langage inclut toutes les structures OWL avec certaines restrictions, comme la séparation des types : une classe ne peut pas aussi être un individu ou une propriété. Il est nommé DL car il correspond à la logique descriptive.

- Le langage OWL Full se destine aux personnes souhaitant une expressivité maximale, ainsi que la liberté syntaxique de RDF, mais sans garantie de calcul. En OWL Full, une classe peut également être un individu, il n'y a pas de séparation des types. C'est pour cela que la calculabilité ne peut être garantie.

A terme, OWL est amené à voir son expressivité augmenter, et devra devenir un véritable langage opérationnel pour permettre l'émergence d'un Web sémantique conforme aux ambitions proposées par T. Berners-Lee (Furst, 2004).

OWL permet de définir :

- Les classes : Les classes fournissent un mécanisme d'abstraction permettant de grouper des ressources partageant des caractéristiques similaires. Comme dans RDF, chaque classe OWL est associée à un ensemble d'individus appelés extension de la classe. Chaque individu de l'extension de la classe est une instance de la classe. Cependant, la notion de classe de OWL a pour but d'être plus expressive, elle a donc été redéfinie (Cardoner, 2004).

- Les propriétés : Une propriété est une relation binaire, s'appliquant soit entre une instance de classe et un littéral ou un type XML (premier type) ou soit entre les instances de deux classes (second type). Contrairement à RDFS, OWL permet de faire la différence entre ces deux cas de figures : DatatypeProperty permet de définir une propriété du premier type alors que ObjectProperty définit le second (Cardoner, 2004).

1.2.3.4 Langages de description et de composition de services :

On va présenter dans cette partie objectifs et fonctionnalités de deux des langages et standards concernant les services sur le web (ou web services). UDDI et WSDL.

1.2.3.4.1 UDDI (Universal Description, Discovery and Integration) :

Le protocole UDDI (Universal Description, Discovery and Integration) est une plateforme destinée à stocker les descriptions des services web disponibles, à la manière d'un annuaire de style «Pages Jaunes». Des recherches sur les services peuvent être effectuées à l'aide d'un système de mots-clés fournis par les organismes proposant les services. UDDI propose également un système de «Pages Blanches» (adresses, numéros de téléphone, identifiants…) permettant d'obtenir les coordonnées de ces organismes. Un troisième service, les «Pages Vertes», permet d'obtenir des informations techniques détaillées à propos des services et permettent de décrire comment interagir avec les services en pointant par la suite vers un PIP RosettaNet ou une "service interface" WSDL. Le vocabulaire utilisé pour les descriptions obéit à une taxonomie bien précise afin de permettre une meilleure catégorisation des services et des organismes.

De par sa simplicité, UDDI permet de stocker l'ensemble des services web sur un seul serveur, dont le contenu est dupliqué et synchronisé sur plusieurs sites miroirs. Des implémentations d'UDDI ont été réalisées, et on peut d'ores et déjà enregistrer son entreprise et les services proposés sur UDDI. Cependant, on peut s'interroger sur la réelle efficacité en matière de recherche d'une architecture aussi simple où la sémantique des données est inexistante et où la description des services se limite à des mots-clés sur lesquels aucune approximation n'est possible. De plus, il n'est pas certain que des serveurs uniques puissent supporter la charge du nombre de services à venir (Baget et al., 2003).

1.2.3.4.2 WSDL :

WSDL est un langage basé sur XML servant à décrire les interfaces des services web, c'est-à-dire en représentant de manière abstraite les opérations que les services peuvent réaliser, et cela indépendamment de l'implémentation qui en a été faite. Il ne comporte pas de moyen de décrire de manière plus abstraite les services (tâche plutôt dévolue à DAML-S ou à UDDI), ni de moyen de conversation et de transaction de messages (tel que SOAP ou d'autres implémentations spécifiques), mais est en général utilisé comme passerelle entre ces représentations de haut niveau et de bas niveau (Baget et al., 2003).

Une description WSDL d'un service contient (Phan, 2005):

- Les types de paramètres d'entrée et de sortie (XML Schéma)
- Les messages échangés entre le service et ses clients
- L'ensemble des opérations fournies par le service (PortType). Les opérations sont les messages d'entrée et de sortie échangés entre le client et le service pour effectuer une opération.
- Les protocoles et les formats de données "concrets" pour les opérations d'un type de port particulier (binding).
- La définition du service Web est considérée comme un ensemble de ports. Elle a aussi les points d'accès uniques définis sous forme d'un binding et d'une adresse du réseau (port).

1.2.4 Architecture en couches du web sémantique :

L'évolution des travaux réalisés dans le cadre du web sémantique est marquée par différents niveaux de complexité. Si les plus élémentaires consistent en une simple affectation de métadonnées aux ressources utilisées, d'autres se basent sur des inférences très alambiquées et permettent l'exploitation de ressources hétérogènes. Quel que soit le niveau de complexité, ces applications reposent toutes sur une architecture en couches commune exprimée par la figure 4, recommandée par le W3C. Il s'agit de la vision schématique du web sémantique proposée par son inventeur Tim Berners-Lee en 2001 (Mestiri, 2007).

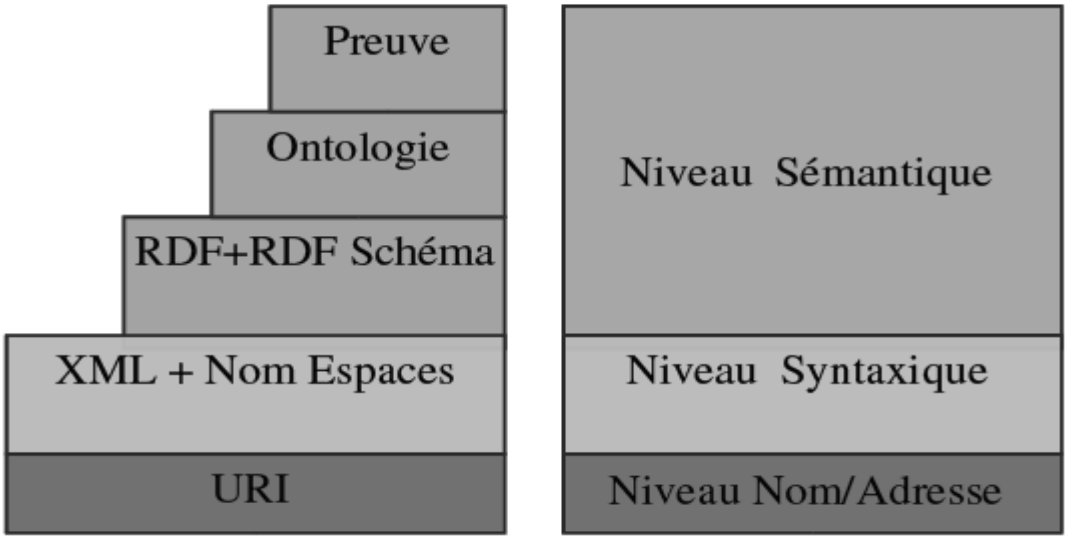

Figure 4. : Architecture en couches du web sémantique.

Deux types de bénéfices peuvent être attendus de cette organisation mentionnée dans (Laublet et al., 2002) :

- Elle permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs.
- Par ailleurs, si elle est bien conçue, elle doit permettre de disposer du langage au bon niveau de complexité, celle-ci étant fonction de l'application à réaliser.

Nous allons maintenant présenter les trois niveaux importants du Web Sémantique : Le niveau de l'adressage, le niveau syntaxique et le niveau sémantique (Phan, 2005).

1.2.4.1 Niveau «Nommage/Adressage » :

Le World Wide Web repose sur un concept important qu'est l'URI (Uniform Resource

Identifier). Tout ce qui est disponible sur Internet doit être identifié par un URI. Un URI identifie de manière unique et non ambiguë chaque ressource du Web, comme une page, une adresse email, ou une image. Le point central des URIs est l'URL (Uniform Resource Locator) traditionnelle utilisée pour définir les liens du Web (par exemple : http : //www.avunet.info.free.fr). Ces URLs sont utilisées pour référencer des fichiers Web à travers un protocole particulier, comme HTTP ou FTP.

1.2.4.2 Niveau Syntaxique :

Le niveau syntaxique est le niveau de la structuration des documents. La spécification de la structure logique des documents repose sur XML (eXtensible Markup Language). XML est déjà présenté dans les langages de web sémantique.

1.2.4.3 Niveau Sémantique :

RDF (Resource Description Framework) est un standard permettant la mise en place de descriptions simples. XML est à la syntaxe, ce que RDF est à la sémantique. RDF Schéma permet ensuite de combiner ces descriptions en un seul vocabulaire. À tout ceci, il manque la possibilité de décrire des vocabulaires spécifiques à des domaines bien particuliers. C'est là que les ontologies jouent leur rôle.

Une ontologie définit les termes utilisés pour décrire et représenter un champ d'expertise. Les ontologies sont utilisées par les personnes, les bases de données, et les applications qui ont besoin de partager des informations relatives à un domaine bien spécifique, comme la médecine, la fabrication d'outils, l'immobilier, la réparation d'automobiles, la gestion de finances, etc..

1.2.5 Applications du web sémantique :

Parmi les diverses applications du web sémantique, On va citer le ecommerce et le elearning.

1.2.5.1 Ecommerce

Le ecommerce va nous permettre d'avoir un échange plus fluide d'information et de transactions entre tous les acteurs économiques, depuis l'offreur de produits ou les services jusqu'aux les clients. Les applications du B2B ont une plus longue histoire et utilisent les échanges informatisés via des structures de messages et de protocoles très codifiées, préétablies et normalisées. Jusqu'à aujourd'hui, la plupart des B2B transactions est basée sur EDI (Electronic Data Interchange ou Échange de Données Informatisés) qui récemment assouplies via des standards basés sur XML (eXtensible Markup Language). Cependant, cette approche semble qu'il sera saturé dans un proche l'avenir. Le commerce électronique Internetbasé fournit un niveau beaucoup plus élevé

de flexibilité et de franchise qui aidera à optimiser des rapports d'affaires. Au lieu de mettre en application un lien à chaque fournisseur, un fournisseur est lié à un grand nombre de clients potentiels. Par conséquent, un fournisseur ou un client peut choisir entre un grand nombre de clients potentiels et peut optimiser ses rapports d'affaires. Donc, on a besoin d'une autre approche : PairàPair (P2P). C'est-à-dire : quiconque doit pouvoir commercer et être en pourparlers avec quiconque d'autre. Cependant, un commerce électronique si ouvert et flexible doit traiter beaucoup d'obstacles avant que ce devienne réalité. Par exemple:

– *Comment peut on trouver et comparer entre des fournisseurs et leurs offres?* Actuellement, presque tout ce travail est effectué manuellement qui apporte des entraves sérieusement à la capacité scalaire du ecommerce.

Donc, la technologie Web sémantique peut lui faire une façon différente pour mécaniser ces tâches parce que la machine peut traiter la sémantique de l'information.

– *Comment peut on faire face à de nombreux et hétérogènes formats de données ? Et à quel «standard» parmi les diverses choses on doit utiliser pour décrire des produits et des services, des catalogues de produit, et des documents d'affaires ?* Donc, la technologie d'Ontologie va nous aider de définir de telles normes mieux et de faire la correspondance entre elles.

– *Et encore, comment peut on faire face à de nombreux et hétérogènes logiques d'affaires ?* On sait que les diverses normes (ou «standards») existent, qui définissent la logique d'affaires d'un partenaire commercial. Donc, la médiation est nécessaire pour compenser ces différences, permettant à des associés de coopérer correctement.

Voilà, l'application de la technologie Web sémantique pour apporter le ecommerce à sa pleine capacité est une activité si prometteuse parce qu'elle peut résoudre efficacement certains obstacles principaux dans ce domaine (Phan, 2005).

1.2.5.2 ELearning

On sait que l'objectif du eLearning est de remplacer les anciennes façons concernant le temps, la place, le contenu de l'apprentissage prédéterminé avec des processus d'apprentissage à temps, à la place de travail , de manière personnalités et à la demande de l'utilisateur (Phan, 2005).

La propriété clé de l'architecture du Web Sémantique (sens partagé commun, métadonnées traitables par les machines), offerte par un ensemble adéquat d'agents, apparaît suffisamment puissante pour satisfaire les exigences du système du e-Learning: rapide, juste à temps et apprentissage pertinent que nous avons illustrés en introduction. Le matériel e-Learning est sémantiquement annoté et pour de nouvelles demandes, il peut être facilement combiné en un nouveau cours d'apprentissage. Selon ses préférences, un utilisateur peut facilement trouver le contenu d'apprentissage utile. Le processus est basé sur les requêtes Web Sémantique et la navigation à travers le matériel d'apprentissage activée par un background ontologique (Boutemedjet, 2004).

Avec le Web sémantique, on peut avoir une plateforme adéquate pour implémenter un système eLearning. On a besoin les moyens pour développer d'une ontologie d'apprentissage. Le matériel d'apprentissage va être annoté par ontologie, et puis, on doit leur composer dans des cours et faire la livraison active des cours à travers des portails d'apprentissage (Boutemedjet, 2004).

Dans le tableau suivant, nous présentons les possibles utilisations du Web Sémantique pour la réalisation des exigences du e-Learning.

Exigences	e-learning	Web sémantique
Livraison	Etudiant détermine son agenda	Les matériaux d`étude sont distribués sur le Web mais eux sont liés a une ontologie. Donc, on peut construire d`un cours spécifique pour l`utilisateur grâce aux requêtes sémantique pour des matières d`intérêt.
Réponse	La réponse réaction au problème actuel	Les propres agents personnalités sur le Web sémantique peuvent employer la langue de service généralement convenue, qui permet la coopération entre ces agents et la livraison active des matériaux d`étude dans le contexte des problèmes actuels.
Accès	Non linéaire Permet l'accès direct à la connaissance dans n'importe quelle ordre pour avoir le sens de la situation actuelle.	l`utilisateur peut décrire la situation actuelle (le but de l`étude, la connaissance précédente..) et exécuter la demande sémantique pour le matériel les plus approprie. Le profil d`utilisateur est également explique. L'accès à la connaissance peut être augmenté par la navigation sémantiquement.
Personnalisation	Le contenu personnalisé est déterminé par les besoins de l'utilisateur et vise à satisfaire aux besoins de chaque utilisateur.	L'utilisateur ayant son agent recherche le matériel d'étude adapté aux besoins de celui. L'ontologie est le lien entre les besoins d'utilisateurs et les caractéristiques du matériel d'étude.
Adaptation	Les changements du contenu dynamique basés sur l'entrée d'utilisateur, ses expériences, ses nouvelles pratiques, les règles d'affaire et l'heuristique quand même.	Web sémantique permet d'utiliser la connaissance fournie dans diverses formes, par l'annotation sémantique de contenu. La distribuée du web sémantique permet d'améliorer des études continues.

Tableau 1. : Exigences du eLearning et Web sémantique.

1.3 Ontologie :

«*An ontology is an explicit specification of a conceptualisation » Tom Gruber*

1.3.1 Notion d'ontologie :

Le terme « ontologie », du grec ***οντοσ*** (être) et ***λογοσ*** (science), a été emprunté au domaine de la philosophie dans lequel il signifie « l'essence de l'essentiel». Dans le domaine de la gestion de connaissance, le sens de ce mot est différent. La notion d'ontologie a d'abord été introduite comme « une spécification explicite d'une conceptualisation ». Cette définition a été légèrement modifiée par la suite. Une combinaison des deux définitions peut être résumée ainsi : « *une spécification explicite et formelle d'une conceptualisation partagée* » (Cardoner, 2004).

Cette définition s'explique ainsi :

- *explicite* signifie que le « type des concepts et les contraintes sur leurs utilisations sont explicitement définies ».
- *formelle* se réfère au fait que la spécification doit être lisible par une machine.
- *partagée* se rapporte à la notion selon laquelle une ontologie « capture la connaissance consensuelle, qui n'est pas propre à un individu mais validée par un groupe ».
- *conceptualisation* se réfère à « un modèle abstrait d'un certain phénomène du monde basé sur l'identification des concepts pertinents de ce phénomène ».

De nombreux types de structure de connaissance se cachent derrière le mot ontologie (taxonomie, thesaurus,…). Ces structures de données peuvent être à la fois terminologiques et conceptuelles. Cependant, ces structures peuvent différer par leur contenu (connaissances générales ou connaissances d'un domaine), par le type des relations sémantiques entre les concepts (relations taxonomiques, relations métonymiques, …), elles se cantonnent par exemple à décrire des liens sémantiques du type "est une sorte de" et son inverse "est représenté par" ou, plus spécifiquement, "est une sous-classe de." Et par le niveau de formalisation (représentation logique,

représentation dans un langage dédié aux ontologies…) (Cardoner, 2004 ; Ghafour, 2003).

Les applications informatiques liées aux ontologies sont multiples: gestion des connaissances, traitement du langage naturel, commerce électronique, etc (Benoit, 2007).

1.3.2 Rôle de l'ontologie :

Nous pouvons énumérer un certain nombre d'utilités d'ontologies, notamment :

- Les ontologies permettent la modélisation des connaissances dans un domaine particulier, dans lequel opère le système à développer (Mestiri, 2007). Une ontologie peut être utilisée comme un répertoire dans lequel on stocke et organise des connaissances et des informations. Elle peut concerner des données simples, standardisées dans un domaine particulier ou bien des données distribuées (Ranwez, 2000).

- Les ontologies assurent une communication fiable et hétérogène entre personnes et machines (agents logiciels ou organisations) du fait qu'elle permet de mettre en place un langage ou un vocabulaire conceptuel commun (Mestiri, 2007), Alors, Une ontologie peut être utilisée comme la base d'un langage de représentation des connaissances (Ranwez, 2000).

- La représentation explicite des connaissances dans un domaine donné sous forme d'une ontologie, permet à son tour une plus grande réutilisation, un partage plus large et une interopérabilité plus étendue (Mestiri, 2007). En 91, Thomas Gruber insistait sur le rôle que pouvaient tenir les ontologies pour favoriser la modularité et la réutilisabilité dans les systèmes informatiques (Gruber, 1991).

Par exemple: On a certain nombre de sites Web contiennent de l'information médicale ou fournissent des services de ecommerce en médecine. Si ces sites partagent et publient tous la même ontologie, qui est à la base des termes qu'ils utilisent, alors les agents informatiques peuvent extraire et agréger l'information de ces différents sites. Les agents peuvent utiliser cette information agrégée pour pouvoir répondre aux

interrogations des utilisateurs ou comme données d'entrées pour d'autres applications (Phan, 2005).

- La représentation conceptuelle des éléments du domaine, permet aux systèmes de réaliser des raisonnements logiques, qu'on appelle inférences, et de sortir avec des conclusions capables d'aider l'utilisateur ou le gestionnaire dans ses décisions (Mestiri, 2007).

- L'indexation et la recherche d'information **:** Dans le web sémantique, d'une façon générale, et dans notre application en particulier, les ontologies sont utilisées pour indexer et décrire les ressources utilisées. Cela permet une plus grande précision dans les résultats des recherches (Mestiri, 2007).

1.3.3 Les Composants d'une ontologie :

L'ontologie n'est en fin de compte qu'une modélisation du monde réel en concept et relation entre ces concepts (Mestiri, 2007). Selon Gruber, la formalisation d'une ontologie se met en place grâce à 5 types de composants (« Modelling primitives ») (Gaelle, 2002). Les principales composantes qu'on peut distinguer sont donc les suivantes :

1.3.3.1 Les concepts

C'est la représentation abstraite des éléments du domaine. On peut également les appeler termes ou classes (Mestiri, 2007). Une classe ou un concept, représente un type d'objet dans l'univers. Les classes sont habituellement organisées en taxinomies auxquelles on applique des mécanismes d'héritage. Tous les concepts peuvent être organisés en une large taxinomie. Il peut aussi y avoir un grand nombre de hiérarchisations plus petites, ou bien pas de taxinomie explicite du tout (Gaelle, 2002).

Les concepts peuvent être abstraits ou concrets, élémentaires (électrons) ou composés (atome), réels ou fictifs. Il arrive que les définitions des ontologies aient été diluées, en ce sens que les taxinomies sont considérées comme des ontologies complètes.

1.3.3.2 Les relations

Les relations représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles (Gaelle, 2002). Les différents types de relations qui peuvent exister sont : « *Spécialisation/Généralisation* », « *Agrégation ou Composition* », « *associé à* », « *composé de* », etc (Mestiri, 2007).

1.3.3.3 Les fonctions

Les fonctions sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini à partir des n-premiers. Comme exemple de fonctions binaires il y a la fonction mère-de ou carré-de, comme fonction ternaire, le prix d'une voiture usagée sur lequel on peut se baser pour calculer le prix d'une voiture d'occasion en fonction de son modèle, de sa date de construction et de son kilométrage (Gaelle, 2002).

1.3.3.4 Les axiomes

Les axiomes sont utiles à la structuration de phrases qui sont toujours vraies. Ils permettent de contraindre les valeurs de classes ou d'instances.

1.3.3.5 Les instances

Ce sont des exemples particuliers de concepts (Mestiri, 2007), ils sont utilisés pour représenter des éléments dans un domaine (Gaelle, 2002).

Nous en présenterons un composant en plus ci-dessous, de Sowa, le rôle.

1.3.3.6 Les rôles

Selon Sowa, « un rôle caractérise une entité par quelque rôle qu'elle joue dans sa relation à une autre entité. Le type « Humain », par exemple, est un type de phénomène qui dépend de la forme interne de l'entité ; mais la même entité peut être caractérisée par des rôles du type, Mère, Employé ou Piéton. » (Gaelle, 2002).

1.3.4 Les types d'ontologie :

Les types d'ontologies mises au point sont très diverses. On présente comme dans (Gaelle, 2002) les types d'ontologies les plus couramment utilisés. D'une façon générale,

on identifie les catégories suivantes: les ontologies de haut-niveau et les ontologies spécialisées, constituées principalement des ontologies de domaine, d'application et de tâches.

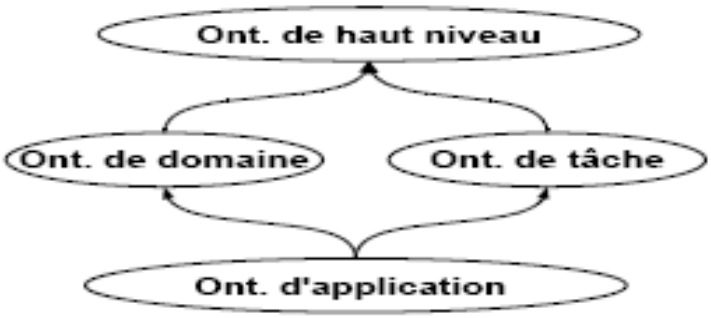

Figure 5. : Les types d'ontologies.

1.3.4.1 Les ontologies de haut-niveau

Ce type d'ontologie décrit des concepts très généraux ou des connaissances de sens commun telles que l'espace, le temps, l'événement, l'action…, qui sont indépendantes d'un problème ou d'un domaine particulier. Qu'elles soient appelées « top level ontologies », « ontologies de sens commun/ général » ou encore « meta-ontologies » ou « ontologies génériques ou noyaux d'ontologies », ces ontologies de haut-niveau (Top-Level ontology, ou Upper Level ontology) fournissent des notions générales auxquelles tous les termes des ontologies existantes doivent être reliés. Elles sont réutilisables d'un domaine à l'autre (définies en relation de « partie-de » et ses propriétés). Une ontologie de haut niveau est généralement conçue afin de réduire les incohérences des termes définis plus bas dans la hiérarchie. Elles incluent du vocabulaire en lien avec les choses, les événements, le temps, l'espace, la cause, le comportement, etc (Gaelle, 2002).

Parmi les ontologies générales il y a (Bahloul, 2006):

- Cyc: développée avec le modèle logique, en utilisant le langage CycL. Cette ontologie a la possibilité de construire des applications pour l'extraction des connaissances, la recherche intelligente et la traduction, etc.
- KR Ontologie : qui utilise le modèle de treillis et le FAC (Formal Concept Analysis) pour représenter l'ontologie.

1.3.4.2 Les ontologies spécialisées

Ce sont des ontologies qui « spécialisent » un sous-ensemble d'ontologies génériques en un domaine ou un sous-domaine. Elles peuvent être de domaine, d'application, techniques (Gaelle, 2002). Les trois principales sont :

1.3.4.2.1 Les ontologies de domaine :

Ce type d'ontologies exprime des conceptualisations spécifiques à des ontologies de domaines particuliers (Bahloul, 2006). Ce sont des ontologies réutilisables au sein d'un domaine donné, mais pas d'un domaine à un autre (Gaelle, 2002).

Ce type d'ontologie décrit un vocabulaire en relation avec un domaine générique comme la médecine ou la physique. Elles se retrouvent dans la typologie de (Heijst et al., 1997) grâce à une classification selon le sujet de conceptualisation.

De nombreuses ontologies de domaine ont été développées notamment dans le domaine de modélisation de l'entreprise (Bahloul, 2006): Entreprise Ontology, Tove, et dans le domaine médical UMLS.

1.3.4.2.2 Les ontologies d'application :

Les ontologies d'application sont généralement spécifiques à une application ; Elles contiennent suffisamment de connaissances pour structurer un domaine particulier. Selon Guarino, ce type d'ontologie décrit des concepts qui dépendent à la fois d'un domaine particulier et d'une tâche particulière. Elles seraient souvent des spécialisations à la fois des ontologies de domaine et des ontologies de tâches et correspondraient aux rôles joués par les entités de domaine lorsqu'elles effectuent certaines activités (Guarino, 1998 ; Gaelle, 2002).

1.3.4.2.3 Les ontologies de tâche :

Selon Guarino, ce type d'ontologie décrit un vocabulaire en relation avec une tâche ou une activité générique comme le diagnostic ou la vente. Les ontologies de tâche fournissent un lexique systématisé de termes utilisés pour résoudre les problèmes associés à des tâches particulières (dépendantes ou non du domaine). Ces ontologies

fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (par ex., plan, objectif, contrainte), des verbes génériques (par ex., assigner, classer, sélectionner), des adjectifs génériques (par ex., assigné) et d'autres mots qui relèvent de l'établissement d'échéances (Gaelle, 2002 ; Guarino, 1998).

1.3.5 La Construction d'une ontologie :

1.3.5.1 Principes de construction d'une ontologie :

Plusieurs travaux se sont intéressés à l'élaboration de principes de construction d'ontologies. Gomez a énuméré un certain nombre de principes à suivre pour l'élaboration d'une ontologie, inspirés par les différents travaux existants (Gruber, 1993 ; Abrouk, 2006):

- Clarté et objectivité : des définitions objectives des termes doivent être fournies afin de clarifier le sens des termes ;
- Exhaustivité : une définition exprimée par une condition nécessaire et suffisante est préférable à une définition exprimée seulement par une condition nécessaire ou seulement par une condition suffisante ;
- Cohérence : une ontologie doit être cohérente afin de formuler des inférences cohérentes avec les définitions ;
- Extensibilité : l'enrichissement de l'ontologie ne doit pas influer sur les définitions existantes ;
- Interventions ontologiques minimales : une ontologie doit faire un minimum d'hypothéses sur le monde en phase de modélisation ;
- Distinction ontologique : les classes de l'ontologie doivent être séparées ;
- Minimisation des distances sémantiques entre les concepts frères : les concepts frères doivent être proches sémantiquement.

1.3.5.2 Le cycle de vie d'une ontologie :

Le cycle de vie d'une ontologie est composé des étapes suivantes (Abrouk, 2006):

- Évaluation des besoins ;
- Construction ;
- Diffusion ;
- Utilisation.

Avant de construire une ontologie, il faut définir son domaine, son utilité et son maintien.

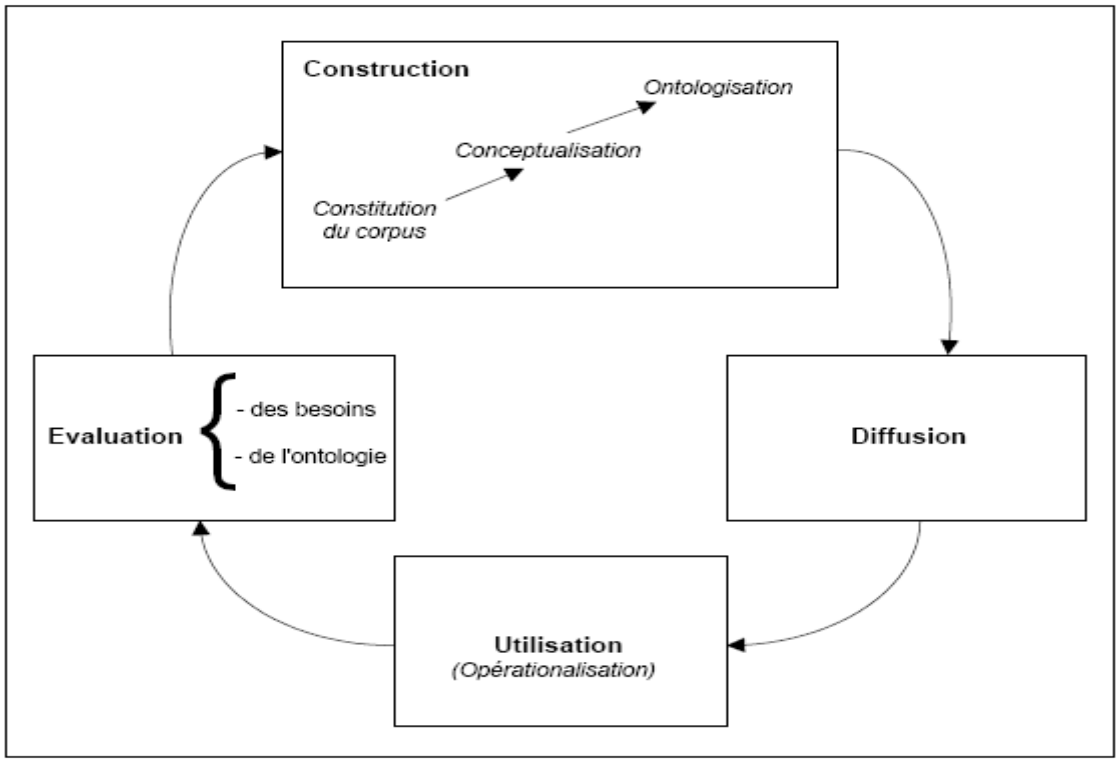

Figure 6. : Le cycle de vie d'une ontologie.

1.3.5.3 Les méthodologies de construction d'ontologies

L'ingénierie ontologique ne propose à l'heure actuelle, aucune méthode normalisée ou méthodologie générale de construction d'ontologies, ce qui rend le processus d'élaboration des ontologies long et coûteux. Ces méthodologies peuvent porter sur l'ensemble du processus et guider l'ontologiste (L'ontologiste est celui qui construit des ontologies. Son travail touche, d'une part, à l'informatique, à la logique, aux modèles de représentation de connaissances, aux normes et standards dans ce domaine et, d'autre part, à la linguistique et aux sciences cognitives) à toutes les étapes. La phase de construction d'une ontologie opérationnelle suivant Frédéric FÜRST peut être

décomposée en trois étapes qui sont la conceptualisation, l'ontologisation et l'opérationnalisation (Furst, 2004) la figure illustre ces étapes.

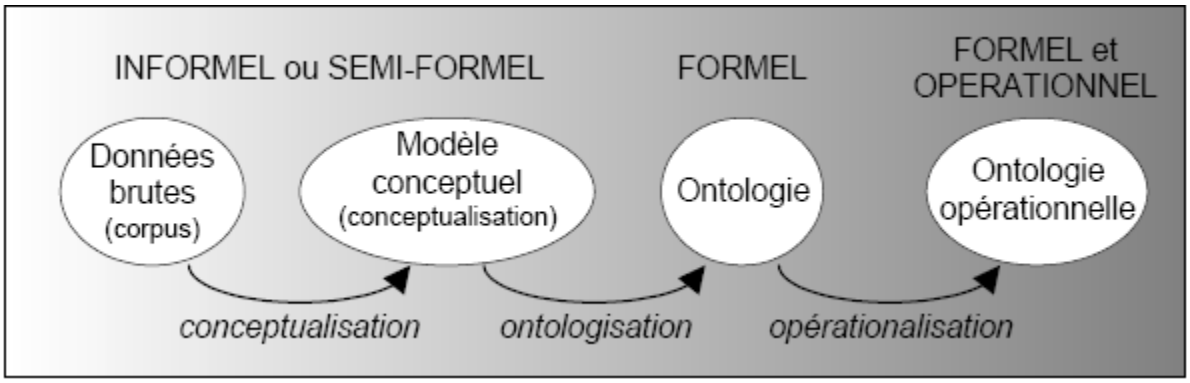

Figure 7. : Construction d'une ontologie opérationnelle.

Mais quelque 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, voire les futurs utilisateurs de l'ontologie. Dans tous les cas, l'élaboration d'une ontologie est basée sur des ressources linguistiques et cognitives constituant un corpus (Un corpus peut être constitué par des interviews d'experts du domaine, des documentations techniques, etc.).

1.3.5.3.1 La constitution d'un corpus :

La construction d'une ontologie suppose au minimum que soit délimité aussi précisément que possible le domaine de connaissance à modéliser, au besoin en le découpant en termes de connaissances du domaine, connaissances de raisonnement, connaissances de haut niveau (communes à plusieurs domaines). Délimiter un domaine repose sur l'utilisation de ressources textuelles et/ou multimédia, constituant le corpus du domaine, et au travers desquelles peuvent être appréhendés la terminologie du domaine et les significations des concepts (Furst, 2004).

Biébow dit : « *Le point de vue actuel est qu'un domaine n'a de sens que défini par un corpus et une application en vue de laquelle l'étude terminologique est effectuée* ».

En effet, un domaine n'est pas seulement défini par le champ de connaissance qu'il couvre, mais aussi par le point de vue sous lequel les utilisateurs de l'ontologie considèrent ce champ de connaissance (Uschold et King, 1995).

Une fois le corpus constitué, la phase de conceptualisation du processus de construction de l'ontologie peut débuter.

1.3.5.3.2 La conceptualisation :

La conceptualisation est un processus d'abstraction qui consiste à identifier les concepts essentiels du domaine de connaissances et d'établir les relations entre ces concepts, à partir d'un corpus représentatif de domaine (Bahloul, 2006). En outre, s'il est prévu d'intégrer d'autres ontologies, les connaissances spécifiées dans ces ontologies ne doivent pas être prises en compte. La nature conceptuelle (concepts, relations, propriétés des concepts et relations, axiomes) des connaissances ainsi extraites du corpus doit ensuite être précisée. Des choix liés aux contextes d'usage de l'ontologie doivent donc être effectués dès cette étape.

La découverte des connaissances d'un domaine peut s'appuyer à la fois sur l'analyse de documents et sur l'interview d'experts du domaine. L'analyse informelle des textes doit être doublée par une analyse automatique qui permet de détecter les termes et structures sémantiques (définitions, règles) présentes dans le corpus (Furst, 2004).

Néanmoins, cette analyse de corpus ne peut suffire à elle seule à spécifier la sémantique du domaine. Car certaines connaissances qui y sont représentées ne prennent sens que lorsqu'elles sont lues par un expert. La sémantique doit donc être précisée ou validée par les experts du domaine (Bahloul, 2006).

Une fois les concepts et relations identifiés par leurs termes, il faut en décrire la sémantique en indiquant, à priori en langage naturel, leurs instances connues, les liens qu'ils entretiennent entre eux, leurs propriétés. La description d'une primitive conceptuelle doit contenir des liens vers les parties du corpus qui mettent en évidence sa sémantique, ce qui permet, au cas où une ambiguïté sémantique demeure, de revenir au corpus. Le processus de conceptualisation mène ainsi à la construction d'un modèle conceptuel (ou conceptualisation), qui décrit les connaissances du domaine. Ce modèle conceptuel n'est cependant pas formel, il peut être complètement informel, exprimé en langage naturel, ou semi-formel, combinant langage naturel et propriétés formelles. Une fois les ressources cognitives passées au travers du tamis de la conceptualisation, il

convient donc, pour l'utiliser dans une machine, de formaliser le modèle conceptuel obtenu. C'est l'objet du processus d'ontologisation (Furst, 2004).

1.3.5.3.3 L'ontologisation :

L'ontologisation consiste à structurer et formaliser autant que possible la conceptualisation pour construire une ontologie spécifiant la terminologie et la sémantique du domaine à travers un modèle doté d'une sémantique formelle (mais non opérationnelle) (Furst, 2004). C'est-à-dire la réalisée par le biais d'un langage formel ou formalisme qui est un ensemble de composants sémantiques (contenu), de règles structurelles (mode d'emploi) et d'une notation formelle particulière (forme) distinée à organiser les relations entre les éléments constituant l'ontologie (Bahloul, 2006).

1.3.5.3.4 L'opérationalisation :

Décrire les connaissances en terme de concepts, de relations et de propriétés sur ces concepts et relations ne suffit généralement pas pour atteindre l'objectif opérationnel d'un SBC. Il s'agit également de tirer au maximum parti de ce qui fait la spécificité du support informatique par rapport au support écrit traditionnel.

Au niveau opérationnel, les concepts et relations sont définis par les opérations qu'il est possible de leur appliquer. Dans cette optique, nous considérons qu'une représentation de connaissance est opérationnelle si la façon dont cette représentation est utilisée pour raisonner est fixée. Par exemple, une propriété de symétrie d'une relation n'est pas en soi opérationnelle, bien que sa sémantique formelle soit fixée : dire qu'une relation est symétrique ne fixe pas la façon dont cette propriété va être utilisée dans un système opérationnel pour raisonner, car on peut l'utiliser pour, par exemple, inférer des relations ou pour contrôler la validité d'une base de faits.

Nous considérons que la sémantique opérationnelle décrit comment les représentations dotées de cette sémantique opèrent sur d'autres représentations pour atteindre un but (Furst, 2004).

1.3.6 L'évaluation d'une ontologie :

Deux niveaux peuvent être distingués dans l'évaluation d'une ontologie (Furst, 2004):

1.3.6.1 La vérification :

Ce niveau consiste à s'assurer que l'ontologie est conforme à un modèle formel de représentation de connaissances. Cette vérification porte sur des propriétés formelles qui ne peuvent être violées par l'ontologie, sous peine de perdre son expressivité.

On considère que la vérification d'une ontologie repose sur le test de trois grands types de propriétés : la conformité, la cohérence et la minimalité (Furst, 2004).

- La conformité d'une ontologie à un modèle de représentation exprime le fait que les représentations de connaissance incluses dans l'ontologie sont bien conformes au modèle utilisé.

- La cohérence d'une ontologie est déterminée par l'absence de contradictions logiques entre les représentations qu'elle contient. Elle fait appel à la sémantique formelle du modèle de représentation, là où la conformité n'utilise que la syntaxe.

- La minimalité d'une ontologie désigne le fait qu'elle ne contient pas de connaissances superflues, c'est-à-dire des connaissances spécifiées deux fois ou plus ou des connaissances qu'on puisse facilement déduire du reste de l'ontologie.

1.3.6.2 La validation :

Qui consiste à s'assurer de la conformité sémantique de l'ontologie à un domaine de connaissance, c'est-à-dire que la sémantique exprimée dans l'ontologie doit être celle du domaine considéré.

La validation permet de tester la complétude de l'ontologie et la conformité de l'ontologie par rapport au domaine.

- La complétude de l'ontologie par rapport au domaine est assurée si toutes les connaissances du domaine sont présentes dans l'ontologie ;

– La conformité de l'ontologie par rapport au domaine est assurée si les connaissances représentées dans l'ontologie correspondent exactement à la sémantique du domaine.

La validation repose sur l'utilisation de spécifications. Ces spécifications peuvent être celles du comportement attendu du système, mais aussi du comportement interdit (spécifications d'anomalies).

En d'autres termes, la vérification correspond à l'exigence « *building the system right* », relativement à un modèle formel, et la validation correspond à l'exigence « *building the right system* », relativement au domaine de connaissance modélisé. De plus, l'évaluation de l'ontologie en amont de son opérationnalisation est souhaitable pour éviter de propager des erreurs, même si l'opérationnalisation peut être nécessaire pour mener certaines activités d'évaluation. Ainsi, utiliser l'ontologie pour répondre à des questions de compétence nécessite d'avoir opérationnalisé l'ontologie ; le test de la cohérence d'une ontologie peut nécessiter des déductions pour mettre en évidence des contradictions logiques entre axiomes. Cependant, la validité des hiérarchies de concepts et/ou de relations doit être testée dès la phase d'ontologisation, aussi bien du point de vue formel que du point de vue sémantique (Furst, 2004).

1.3.7 L'alignement et la fusion d'ontologie :

Utiliser les connaissances de différents domaines au niveau ontologique passe donc par l'utilisation de plusieurs ontologies couvrant chacune un des domaines visés, domaines a priori non disjoints puisqu'ils sont utilisés au sein du même système d'information. Les représentations considérées sont donc connexes et les utiliser de façon cohérente nécessite de déterminer les parties communes aux différentes ontologies. L'alignement de deux (ou plus) ontologies est nécessaire quand elles interviennent toutes deux dans un même SBC.

1.3.7.1 L'alignement d'ontologies :

Consiste à trouver des correspondances entre les connaissances spécifiées dans les deux ontologies, de manière à pouvoir les exploiter conjointement dans le même système. En pratique, il s'agit d'identifier des concepts (ou des relations) de la première ontologie

avec des concepts (ou des relations) de la seconde, ou de trouver des liens conceptuels (subsomption, ...) entre eux (On parle dans le premier cas d'alignement syntaxique et dans le deuxième d'alignement sémantique (Furst, 2004)). Contrairement à l'alignement où les deux ontologies de départ reste intactes.

En pratique, l'alignement de deux ontologies se fera dans le cadre du même langage de représentation et pourra donc nécessiter des transcriptions préalables d'un langage à l'autre (Furst, 2004).

1.3.7.2 La fusion d'ontologies (merging) :

Consiste, à partir de deux ontologies, à en créer une troisième qui intègre les connaissances spécifiées dans les deux premières.

L'uniformisation des modèles et formalismes de représentation sont également nécessaires à la fusion d'ontologies. Préalablement à la fusion, il convient de déterminer quelle est l'ontologie la plus générale, ou celle qui est la plus étendue. Les autres devront être alignées sémantiquement et syntaxiquement sur l'ontologie la plus générale (Furst, 2004).

Dans les deux cas, la connexité des deux domaines de connaissance modélisés par les ontologies est requise, sans quoi aucun lien ne peut être établi entre concepts. De plus, les formalismes de représentation d'ontologie utilisés doivent être au moins compatibles, ainsi que les paradigmes conceptuels (Maedche et al., 2002).

1.3.8 Outils d'aide à la construction d'ontologie :

De nombreux outils de construction d'ontologies utilisant des formalismes variés et offrant différentes fonctionnalités ont été développés citant :

1.3.8.1 TERMINAE :

TERMINAE, développé au LIPN de l'Université Paris-Nord, permet, à travers l'outil d'ingénierie linguistique LEXTER, d'extraire d'un corpus textuel les candidats termes d'un domaine (Biebow et Szulman, 1999). Ces concepts doivent ensuite être triés par un expert et organisés hiérarchiquement, puis la sémantique du domaine est précisée à travers des axiomes. TERMINAE offre ainsi une aide à la conceptualisation.

1.3.8.2 OntoEdit

OntoEdit (Ontology Editor) (Sure et al., 2002), développé par la compagnie Ontoprise (http://www.ontoprise.de/products/ontoedit), est également un environnement de construction d'ontologies basé sur une méthodologie. Il permet l'édition des hiérarchies de concepts et de relations dans le cadre du paradigme des frames et l'expression d'axiomes algébriques portant sur les relations, et de propriétés telles que la généricité d'un concept. Des outils graphiques dédiés à la visualisation d'ontologies sont inclus dans l'environnement. OntoEdit intègre, dans sa version commerciale, un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs ainsi qu'un plug-in permettant le test de la cohérence d'une ontologie. Enfin, un plugin nommé ONTOKICK offre la possibilité de générer les spécifications de l'ontologie par l'intermédiaire de questions de compétence. OntoEdit gère de nombreux formats de représentation de connaissance dont DAML+OIL, RDFS et FLogic. La figure 8 présente une ontologie des voyages éditée avec OntoEdit (Furst, 2004).

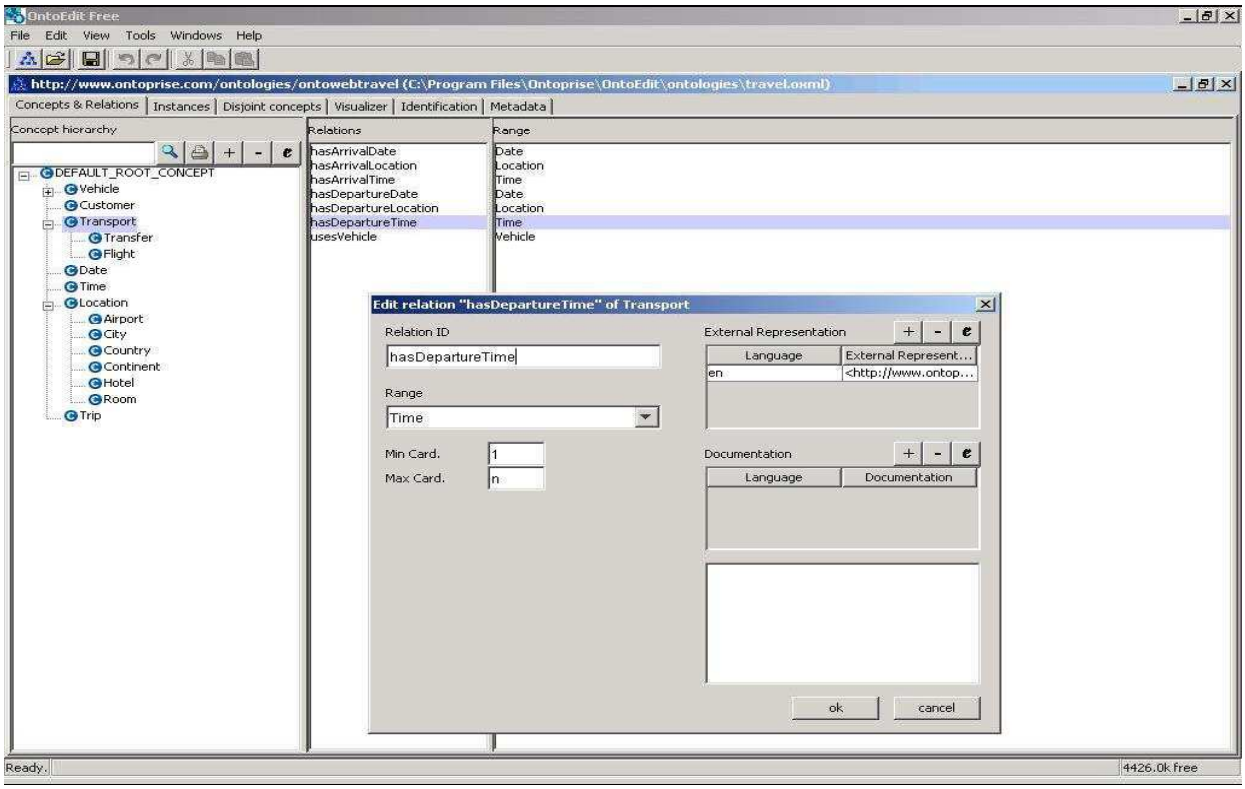

Figure 8. : L'éditeur d'ontologie OntoEdit.

1.3.8.3 L'Editeur d'Ontologies Protégé :

Protégé : est le plus connu et le plus utilisé des éditeurs d'ontologie. Open-source, développé par l'Université de Stanford, il a évolué depuis ses premières versions (Protégé-2000), c'est un éditeur qui permet de construire une ontologie pour un domaine donné, de définir des formulaires d'entrée de données, et d'acquérir des données à l'aide de ces formulaires sous forme d'instances de cette ontologie. Protégé est également une librairie Java qui peut être étendue pour créer de véritables applications à bases de connaissances en utilisant un moteur d'inférence pour raisonner et déduire de nouveaux faits par application de règles d'inférence aux instances de l'ontologie et à l'ontologie elle-même (méta-raisonnement) (Protégé, 2007).

Dans le contexte du web sémantique des « plugin » pour les langages RDF, DAML+OIL et OWL ont été développés pour Protégé. Ces « plugin » permettent d'utiliser Protégé comme éditeur d'ontologies pour ces différents langages, de créer des instances et les sauver dans les formats respectifs.

Il est également possible de raisonner sur les ontologies en utilisant un moteur d'inférence général tel que JESS , ou des outils d'inférence spécifiques au web sémantique basés sur des logiques de description [DL] tels que RACER (Racer : est le moteur d'inférence connu. Il est commercialisé par Racer Systems GmbH & Co. KG, fondé en 2004 par Volker Haarslev, Kay Hidde, Ralf M¨oller et Michael Wessel qui travaillaient `a l'université de Hambourg) . Ces deux outils peuvent être facilement intégrés à Protégé. Les logiques de description permettent de définir les bases logiques des différents formalismes de représentation de la connaissance tant sur le plan de la représentation que sur le raisonnement. Dans les formalismes de représentation de la connaissance, il est souvent nécessaire de restreindre l'expressivité pour rendre certains types de raisonnement, tels que la classification automatique, faisable (« tractable ») (Protégé, 2007). La figure 9 présente Une Ontologie pour le Domaine de l'Immobilier.

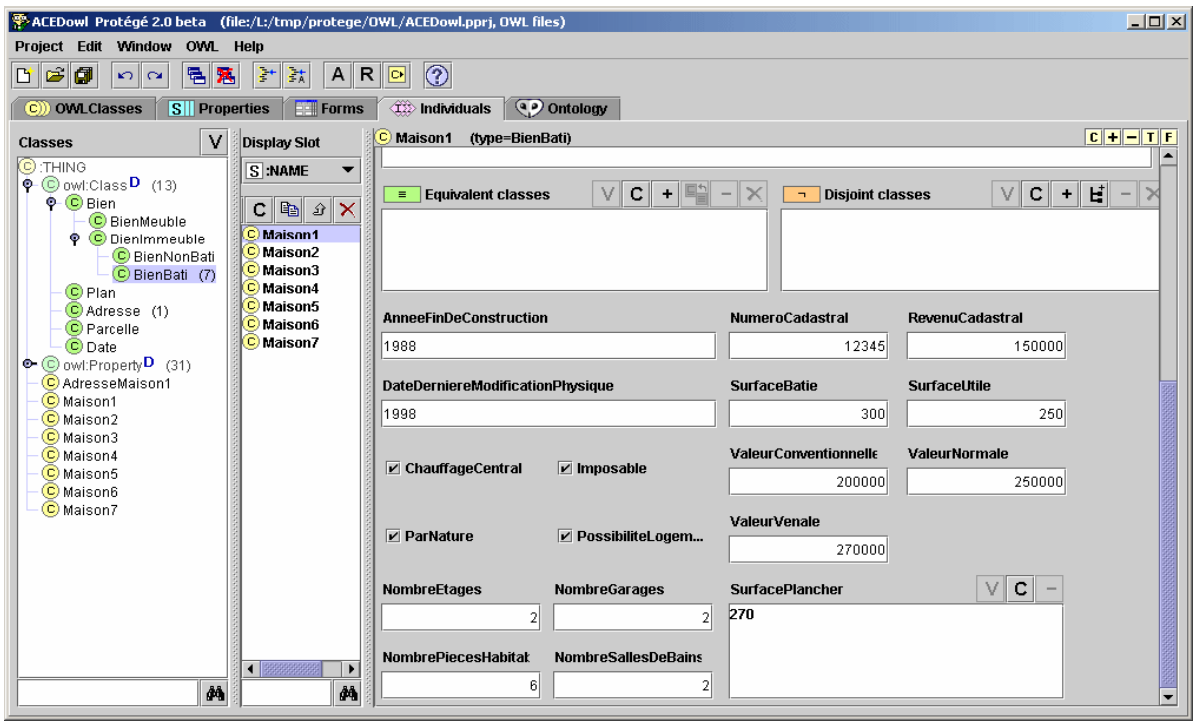

Figure 9. : Editeur d'Ontologie Protégé.

1.3.9 Utilisation des ontologies en Recherche d'Informations :

Les systèmes de recherche d'information visent à restituer (tous) les documents pertinents (et seulement ceux là) par rapport à un besoin d'information exprimé par un utilisateur. Les systèmes d'exploration visent quant à eux à fournir de l'information élaborée à partir de l'analyse d'un ensemble d'informations relatives à un thème. Les index jouent un rôle primordial dans ces deux types de systèmes en définissant les descripteurs (mots ou groupements de mots) qui représentent le contenu des documents et à partir desquels les documents peuvent être accédés ou analysés.

L'indexation des documents n'est pas le seul moyen utilisé pour structurer et organiser une collection. Un modèle de structuration complémentaire consiste à utiliser les métadonnées explicitement associées aux documents. Ces méta-données peuvent être combinées aux descripteurs issus de l'indexation pour fournir une représentation plus complète des documents de la collection. Cette représentation basée sur les métadonnées et les descripteurs correspond à de la connaissance relative au corpus.

Quel que soit le mode de représentation interne des informations par le système, un des problèmes auquel l'utilisateur doit faire face est d'exprimer son besoin en information.

Généralement quand un utilisateur sait ce que contient la collection, comment elle est structurée, ce qu'il recherche et comment il peut le décrire, il n'a pas de problème pour formuler sa requête. En fournissant les méta-données, le système peut aider l'utilisateur à évaluer le potentiel de la collection, cerner ses besoins et définir ses requêtes. Se pose cependant le problème de choisir quelles méta-données fournir à l'utilisateur et quelle organisation choisir pour représenter cette connaissance. Les ontologies semblent être la solution la plus adaptée à cette problématique (Cardoner, 2004).

D'ailleurs, indexer des collections employant un ontology présente les avantages suivants (Hernandez et Mothe, 2007):

- Il aide l'utilisateur à formuler la requête. En présentant l'utilisateur avec l'ontology, il est possible de lui guider dans son choix des termes utilisés dans la requête.
- Il facilite le RI dans les collections hétérogènes en indexant tous les types de documents selon les mêmes concepts.

Dans le contexte IR, un ontology n'est pas habituellement représenté logiquement. Le formalisme utilisé facilite généralement la gestion des concepts comme objets, leur classification, la comparaison de leurs propriétés et navigation dans l'ontology par l'accès d'un concept et ceux reliés avec lui.

1.4 Conclusion :

Le Web sémantique est le Web de demain dans lequel les métadonnées sémantiques jouent un rôle très important. On peut utiliser ce type de données pour enrichir des pages Web existantes. Avec la technologie Ontologie pour la représentation des connaissances, l'annotation des Web et des services pose plusieurs avenues de recherches (Phan, 2005).

Au cours de ce chapitre nous n'avons pas pu tout énumérer à propos du web sémantique et d'ontologies, Nous avons dressé, un rapide état de l'art sur le web sémantique, son interet dans la représentation des connaissances, ses principales technologies et ses applications. Après, on a fait un tour d'horizon autour des

ontologies, en s'appuyant sur leurs types, leurs composants, leur construction et leur alignement et fusion, et enfin on a terminé par son utilité dans la recherche d'informations.

Dans ce qui suit, on va décrire un état de l'art sur la recherche de l'information et l'indexation des documents sur le web.

Chapitre 2 : La recherche d'information et l'indexation documentaire.

2.1 Introduction :

Le développement récent et explosif du World-Wide Web a relancé les travaux sur les outils de recherche d'information textuelle, devenus rapidement indispensables à la navigation sur Internet.

Le nom de « recherche d'information » (information retrieval) fut donné par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maîtrise. La première conférence dédiée à ce thème (International Conference on Scientific Information) s'est tenue en 1958 à Washington. On y comptait les pionniers du domaine, notamment, Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn, etc (Jian-Yun, 2008).

Ce chapitre présente un tour d'horizon autour la recherche d'information, le développement de la langue arabe dans le domaine de la recherche d'information ainsi que l'utilité de l'indexation documentaire dans le cadre de RI.

2.2 La recherche d'informations:

D'après la définition de Salton en 1968, La recherche d'information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information (Boughanem, 2006). Alors, l'objectif de la recherche d'information (RI) est de concevoir des systèmes (nommés désormais SRI pour systèmes de recherche d'information) capables de retrouver parmi un ensemble de

documents ceux qui répondent précisément au besoin d'un utilisateur. Ce besoin est généralement formulé par le biais d'une requête en langage naturel (Moreau, 2006).

2.2.1 Architecture et fonctions d'un système de recherche d'informations:

Un SRI se modélise par le quadruplet SRI=<D, Q, M, P> où :

- D est l'ensemble des documents du corpus ;
- Q est un langage de requête destiné à représenter les besoins d'information de l'utilisateur. Ce langage définit l'ensemble des requêtes que peut formuler directement ou indirectement un utilisateur d'un SRI ;
- M est un modèle de RI qui sert à décrire les documents de D et à exprimer les requêtes de Q.
- P est une fonction qui associe une valeur de pertinence entre toute requête qi de Q et tout document d de D. Cette fonction peut fournir un ordonnancement des documents par rapport à la requête qi (Zargayouna, 2005).

Le SRI se décompose essentiellement en deux processus de base :

- Le processus d'indexation qui consiste à identifier dans un document certains éléments significatifs qui serviront de clés pour retrouver ce document au sein d'une collection (Zargayouna, 2005), alors créer une représentation (index) de leur contenu textuel qui soit exploitable par le SRI (Moreau, 2006), le même processus d'indexation s'applique généralement aux requêtes lors du processus d'interrogation.
- Le processus d'interrogation (ou de recherche) vise à apparier les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Pour cela, elle s'appuie sur un formalisme précis dé ni par un modèle de RI. Les documents présentés en résultat à l'utilisateur, et considérés comme les plus pertinents, sont ceux dont les termes d'indexation sont les plus proches de ceux de la requête (Moreau, 2006).

La recherche des documents à pour objectif de ressortir les documents les plus pertinents de la base pour une bonne interprétation (Bessou et al., 2007).

Cette architecture peut être enrichie par un retour arrière sur pertinence (relevance feedback) qui affine la recherche et améliore la qualité des résultats en tenant compte de l'évaluation de l'utilisateur qui classe les documents en pertinents et non pertinents. Il est aussi possible d'avoir recours à l'expansion de requêtes qui permet d'étendre la requête (en rajoutant des termes) ou la ré-écrire (Zargayouna, 2005).

Les processus d'indexation et de recherche sont dépendants l'un de l'autre comme c'est indique dans la figure suivante :

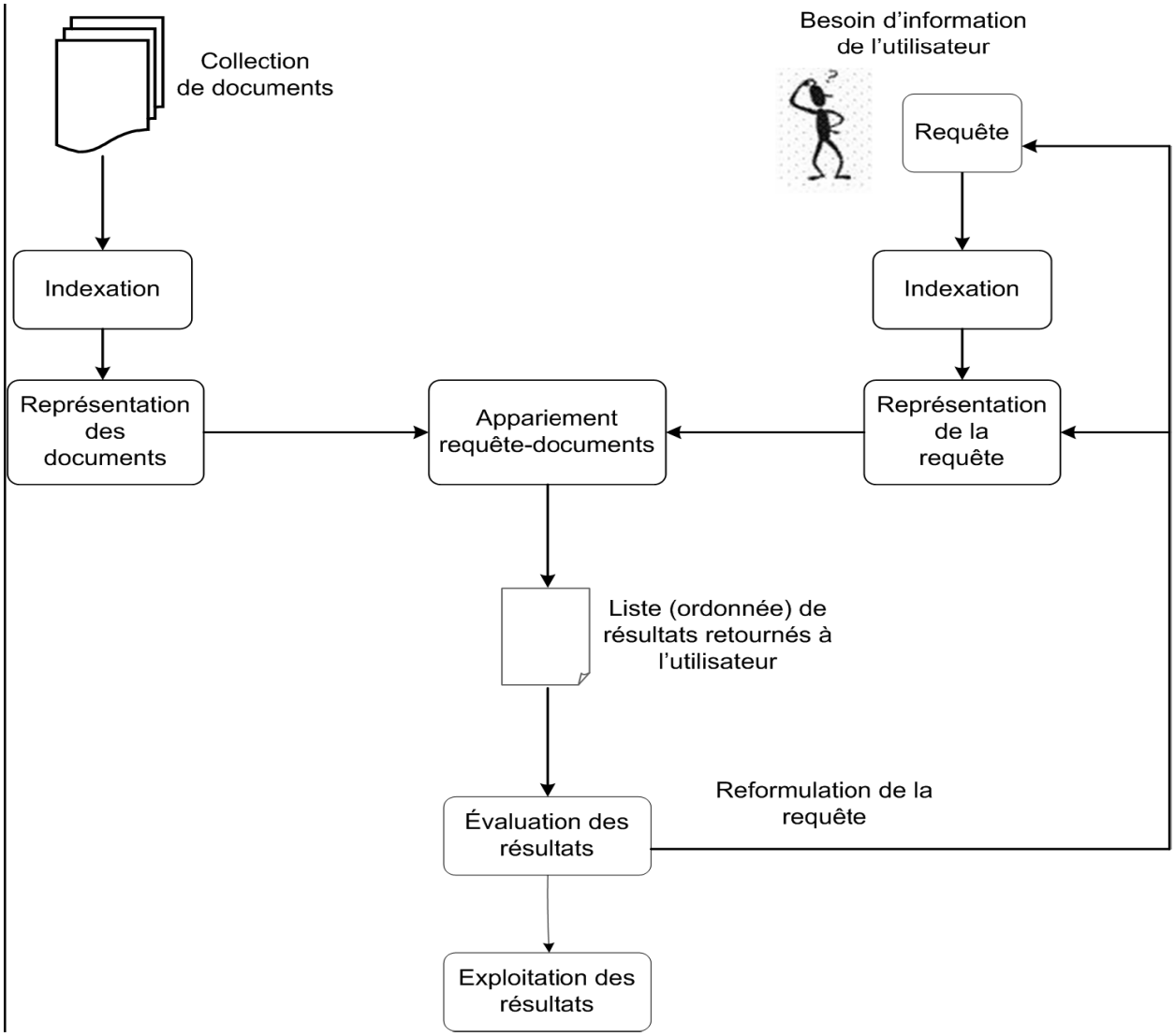

Figure 10. : Processus en U de recherche d'information.

2.2.2 Panorama des outils de recherche d'informations actuels:

L'apparition du World-Wide Web, a conduit à une croissance exponentielle du nombre d'utilisateurs du réseau mais aussi à une croissance exponentielle du nombre de textes accessibles aux utilisateurs. Des quelques centaines de milliers de pages de texte accessibles en 1993, le stock d'information sur Internet atteint aujourd'hui quelques centaines de millions (350 millions en juillet 1998 d'après l'estimation de (Broder et Henzinger, 1998), avec un taux de croissance estimé à 20 millions de pages par mois. Il est donc devenu impossible de naviguer sur cet océan de données et de localiser l'information souhaitée sans des outils appropriés.

C'est ce qui a permis la rapidité de développement d'outils d'aide à la navigation on distingue les annuaires thématiques, qui procèdent à un référencement et une description humaines des sites Web (par exemple la partie annuaire de Yahoo, Nomade, l'Open Directory…) et les moteurs de recherche (Google, Alta Vista, Exalead, Wisenut, YST…), qui fonctionnent par collecte et indexation automatisées des pages Web (et non des sites) (Vignaux, 2007). Les moteurs de recherche apportent une réponse plus " technologique" basée sur des outils informatiques beaucoup plus puissants mais sans intelligence particulière (Bourdoncle, 1999).

Cette distinction, « historique » car elle a longtemps structuré le monde des outils, est moins nette aujourd'hui, à cause de la mixité, de l'imbrication des annuaires et des moteurs : Google utilise l'annuaire de l'Open Directory, Yahoo a son propre moteur, etc (Vignaux, 2007).

Mais le critère des modes d'indexation reste essentiel, car il induit des ressources, des usages et des technologies très différentes. Ainsi un annuaire thématique va-t-il référencer des sites Web, là où un moteur indexera toutes les pages d'un site ; l'annuaire facilitera le défrichage, le premier repérage des ressources dans un domaine ou un secteur défini, par l'organisation arborescente proposée, alors qu'un moteur de recherche permettra de trouver un document très précis. Autrement dit, les deux familles se prêtent à des utilisations complémentaires : pour connaître la liste des journaux présents sur le Web, la navigation dans un annuaire sera recommandée, alors

que vous y trouverez difficilement un support pédagogique sous Power Point, en français, paru en 2002 et traitant du fonctionnement des ordinateurs… (Serres, 2004).

Afin d'améliorer la pertinence des documents retournés, les moteurs de recherche disposent de plusieurs angles d'attaque. L'un des premiers concerne les requêtes elles-mêmes, par exemple la correction orthographique ou la détection automatique des phrases (Bourdoncle, 1999).

Cependant, le principal levier dont dispose l'architecte d'un moteur de recherche reste encore l'amélioration de l'algorithme d'évaluation de pertinence (ranking.) En effet, les algorithmes traditionnels, fondés sur la mise en correspondance des mots des requêtes et des mots contenus dans les documents trouvent rapidement leurs limites sur le World-Wide Web (Bourdoncle, 1999).

Allant faire un tour d'horizon autour des approches de classement des résultats des moteurs de recherche pour aider les utilisateurs dans leurs recherches citons par exemple (Bourdoncle, 1999):

- Une approche permettant de classer les résultats des recherches dans des dossiers thématiques (dont la liste est établie manuellement), a plus récemment été déployée sur le moteur de recherche NorthernLight (www.northernlight.com).
- La fonction *What's Related* de Netscape (www.netscape.com) proposant des liens vers des pages au contenu proche d'une page donnée,
- La fonction *More Like This* du moteur Excite (www.excite.com) permettant d'affiner une requête de manière à rechercher des pages au contenu proche d'un des résultats de cette requête.
- l'approches alternatives de celle suivie par les moteurs de recherche, comme la recherche par nom de marques de *RealNames* (www.realnames.com), la reformulation de requêtes en questions aux réponses connues, voie suivie par AskJeeves (www.askjeeves.com),

- L'approche des *anneaux* (*rings* en anglais) qui consiste à relier entre eux par des liens hypertextes les sites aux contenus voisins (ce qui ne résout toutefois pas le problème de trouver un premier site situé dans l'anneau).

En plus des annuaires et des moteurs de recherche il y a encore les méta-moteurs qui interrogent en parallèle plusieurs moteurs de recherche classiques et fusionnent ensuite de manière intelligente les résultats de ces derniers (Bourdoncle, 1999), les portails et les outils dits annexes. Un portail se distingue notamment des autres outils traditionnels par un ensemble de services personnalisés offerts aux usagers (compte personnel, messagerie, commerce, commande de documents, veille, etc.). Quant aux « outils annexes », il s'agit d'un ensemble d'outils diversifiés, pouvant servir à la recherche d'information et à la veille : « aspirateurs de sites » Web, organisateurs de signets, outils collaboratifs de partage des signets (Vignaux, 2007).

Néanmoins, les moteurs de recherche rendent souvent des centaines de documents pour chaque requête. La tâche la plus lourde revient à l'utilisateur qui doit fouiller dans cette masse d'information pour sélectionner les documents qui lui seront les plus utiles. Les résultats ne sont pas tout pertinents et l'information retrouvée n'est pas complète. Autrement dit, la recherche plein texte n'est pas toujours efficace car il existe des variantes lexicales et des synonymes considérés comme étant des termes différents (Anh, 2005).

La problématique qui se pose est celle d'une recherche d'informations intelligente où l'indexation devrait reposer sur la sémantique des ressources comme étant «l'explication de structures et de concepts contenus dans les documents numériques ou qui leur sont associés». L'intérêt est d'une part d'apporter suffisamment de renseignements sur les ressources, en ajoutant des annotations sous la forme de métadonnées et d'autre part, de décrire leur contenu de manière à la fois formelle et signifiante à l'aide d'une ontologie pour être interprétables aussi bien par les humains que par les machines (Anh, 2005).

2.2.3 Utilisations d'ontologies par les systèmes de recherche d'information:

Les ressources sémantiques (thésaurus, ontologies, etc.) ont un apport considérable pour le traitement des documents textuels ou multimédia. Leur utilisation en Recherche d'Information (RI) peut intervenir lors de la phase de recherche ou lors de la phase d'indexation (Zargaouna, 2005). Les ontologies permettent de mettre en oeuvre des traitements qui assistent les processus de recherche de façon pertinente.

Ces nouvelles formes peuvent être des hiérarchies de concepts ou taxonomies comme dans MeSH, des ontologies de domaine comme dans GO (Gene Ontology), ou des ontologies génériques comme c'est le cas dans PenMan, Cyc et WordNet (Baziz, 2005).

2.2.3.1 Phase d'indexation :

Les documents peuvent être indexés par un groupe de concepts, où on sait qu'un tel document traite des concepts A et B mais on ne connaît pas les relations entre eux dans le texte. Cette méthode permet néanmoins une recherche plus pertinente qu'une recherche par des termes simples, même si les concepts ne sont pas reliés entre eux, les termes référant au même concept le sont. Une autre méthode attribue à chaque document une description sémantique (ou annotation) où les concepts sont représentés avec leurs relations sémantiques. Cette représentation confère un grand pouvoir d'expression mais peut par ce fait ralentir les traitements et la construction des descriptions sémantiques associées à chaque document n'est pas une tâche facile (Zargaouna et Salotti, 2004).

2.2.3.2 Phase de recherche d'information :

L'intérêt d'utiliser des ressources sémantiques en recherche d'information est de pouvoir retourner, lors d'une recherche par similarité, les documents qui partagent avec la requête le maximum de concepts plutôt que le maximum de mots-clés. Les ontologies ont montré leur efficacité en RI, leur utilité s'est vu confirmé par le web sémantique. Une ontologie permet d'affiner les résultats en réduisant le silence et le bruit (Zargaouna et Salotti, 2004).

2.2.4 Tendances de la recherche d'informations sur le web :

Avant de pointer ces évolutions en cours et à venir, il n'est sans doute pas inutile de revenir brièvement en arrière et de mesurer le chemin parcouru ; le regard rétrospectif, le souci de l'histoire, toujours nécessaire, sont de plus en plus oubliés ou minorés aujourd'hui, notamment à cause de l'emballement technologique, où une innovation chasse l'autre tous les six mois, en effaçant les traces des techniques passées. Nous évoquerons quelques-unes des grandes tendances qui peuvent résumer les principaux bouleversements de la recherche d'information depuis les débuts de l'informatisation documentaire, il y a plus d'une quarantaine d'années.

En prenant en compte les différents « composants » de la recherche d'information, Serres dans (Serres, 2004) a relevé sept tendances, sans aucune prétention d'exhaustivité, que l'on peut résumer ainsi :

- de la dépendance à l'autonomie des usagers
- de la maîtrise des stocks à la surabondance des flux
- de la validation *a priori* à la validation *a posteriori*
- de la rareté et de la distinction à l'explosion et à l'hybridation des outils et des modes de recherche
- du « retrouvage » booléen à la « sérendipité »
- du modèle de l'accès à celui du traitement de l'information
- de la gratuité à la commercialisation de la recherche.

2.3 La langue arabe et la recherche d'information sur le web:

La croissance exponentielle du Web a permis à certaines langues telles que l'arabe d'y être présentes. Les statistiques ont montré que, depuis 1995, année de lancement du premier journal électronique arabe sur Internet (asharqulawsat.com), le nombre de sites arabes a augmenté de manière considérable. En l'an 2000, près de 20.000 sites arabes ont été comptabilisés, soit environ 7% des publications disponibles sur Internet (Abdelali et

al., 2004). La recherche d'information en langue arabe est devenue un centre d'intérêt aussi bien pour la recherche que pour le développement commercial et technique. Le nombre d'usagers arabophones du Web recensés en 2002 était de 4,4 millions soit 1,5% de la population arabophone (Levini, 2002). Le problème reste tout de même les insuffisances des moteurs de recherche sur le Web pour l'arabe (El-Hachani, 2005).

2.3.1 Caractéristiques de la langue arabe :

L'application de l'indexation sur la langue arabe pose des problèmes majeurs dont : Le problème de l'ambiguïté issue de l'absence des voyelles, les lettres changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot) ceci exige des règles morphologiques complexes Le problème de reconnaissance des formes fléchies, car l'arabe est une langue fortement flexionnelle, L'arabe s'écrit et se lit de droite à gauche (Bessou et al., 2007 ; Douzidia, 2004).

L'arabe est un des six langues officielles de Royaum Uni. C'est une langue sémantique et la langue maternelle de plus de 150 millions de personnes dans 21 pays arabes (Zaidi et al., 2005). En ce qui concerne les internautes parlant l'arabe, le nombre est estimé à 4,4 millions (Abdelali et al., 2004).

L'alphabet arabe contient 28 caractères Tableau 2 (Douzidaia, 2004). Trois caractères prennent des différentes formes (Zaidi et al., 2005):

- Hamza (ء) est parfois écrite : أ, إ et ا (alif)
- Ta marbouta (ة) comme t en français trouve à la fin sans les deux points (ه = ha)
- Alif maqsurah (ى) est le caractere (ي =ya) sans points. Ces trois caractères posent quelques difficultés dans la réalisation des SRI.

Lettre arabe	Correspondant français	Prononciation
ا	A	Alef
ب	B	Ba’
ت	t	Ta’
ث	th	Tha’
ج	j	Jim
ح	H	Hha’
خ	kh	Kha’
د	D	Dal
ذ	D	Thal
ر	r	Ra
ز	Z	Zayn
س	S	Sin
ش	sh	Shin
ص	S	Sad

Lettre arabe	Correspondant français	Prononciation
ض	d	Dad
ط	t	Tah
ظ	z	Zah
ع	‘ ‘	Ayn
غ	gh	Ghayn
ف	f	Fa
ق	q	Qaf
ك	k	Kaf
ل	l	Lam
م	M	Mim
ن	n	Nun
ه	h	Ha
و	W	Waw
ى	y	Ya

Tableau 2. : Les 28 lettres arabes.

Il y a quelques centres d’information qui ignorent le hamza et les points au dessus de ta marbouta pour unir l’entrée et la sortie pour ces caractères. Il y a en Arabe des séries entières des non-alphabetiques signes, ajoutes au-dessus ou au-dessous des lettres harmonieuses pour rendre la lecture du mot moins ambiguë. Celles-ci s'appellent des voyelles, ou les marques diacritiques (Zaidi et al., 2005). Elles sont nécessaires à la lecture et à la compréhension correcte d’un texte, elles permettent de différencier des mots ayant la même représentation. Le tableau 3 donne un exemple pour les mots **مدرسة et كتب** (Douzidaia, 2004). Cependant, les voyelles ne sont utilisées que pour des textes sacrés et didactiques. Les textes courants rencontrés dans les journaux et les livres n’en comportent habituellement pas (Douzidia, 2004).

Mot sans voyelles	1ère Interprétation		2ème Interprétation		3ème Interprétation	
كتب	كَتَبَ	il a écrit	كُتِبَ	Il a été écrit	كُتُبْ	des livres
مدرسة	مَدْرَسَةٌ	école	مُدَرَّسَةٌ	enseignante	مُدَرَّسَةٌ	enseignée

Tableau 3. : Ambiguïté causée par l’absence de voyelles pour les mots مدرسة etكتب .

L'utilisation de la langue arabe, objet de notre étude dans le domaine d'indexation constitue un grand pas vers son intégration dans la technologie de l'information. Vue sa puissance et sa richesse.

2.3.2 Encodage de la langue arabe :

La langue arabe sur Internet a connu beaucoup de changements lors de sa première introduction dans l'internet en région arabe. Au début, le texte arabe a été affiché en utilisant un fichier GIF ou JPG. Cela puisque les navigateurs Web ne peuvent pas afficher correctement l'arabe. Alors que Microsoft a commencé domine le marché des navigateurs, il a produit ce que l'on appelle (page de code Windows 1256) codant pour la langue arabe. Pourtant, tous les sites Web en arabe sur Windows CP-1256; il y a eu des vérités d'entre eux qui utilisent Unicode (UTF-8) pour l'encodage (Al-Khalifa et Davis, 2005).

La visualisation des textes est plus ou moins difficile selon les langues en raison de la présence de signes diacritiques, de la direction de l'écriture, de la possibilité de ligatures, de la présence ou de l'importance des déclinaisons. L'arabe possède 28 lettres dont les voyelles, longues ou courtes, se placent au-dessus ou au-dessous des consonnes ; ce qui complique leur représentation graphique. La question est comment représenter ces voyelles ?

Représenter chaque caractère vocalisé en tant que code unique n'est pas une solution efficace, il est parfois souhaitable voire possible de visualiser le texte avec ou sans les voyelles. La vocalisation de l'arabe est une problématique importante aussi bien pour la visualisation que pour la recherche d'information (El-Hachani, 2005).

Le second problème est que les systèmes en vigueur ne supportent pas les caractères arabes. En effet, l'existence de certaines ligatures ou liaisons de termes posent des difficultés encore plus complexes, donnons l'exemple du *LAM ALIF*, l'équivalent de l'article défini en arabe, il ne peut s'écrire qu'avec la ligature. Il faut également préciser qu'une lettre en arabe peut prendre deux, trois jusqu'à quatre formes : début, milieu et fin de mot ou isolée (El-Hachani, 2005).

L'autre aspect est le sens de l'écriture, celle de l'arabe est bidirectionnelle, en effet l'arabe s'écrit de droite à gauche alors que le caractère s'écrit de gauche à droite. Les caractères arabes sont stockés dans un ordre logique différent de l'ordre visuel. L'ordre logique correspond à l'ordre de lecture – saisie. L'autre problème pour l'arabe est que très souvent des documents rédigés dans un système (Macintosh par exemple) ne peuvent être visualisés que par quelques navigateurs à cause de l'existence de codes de page différents. C'est le problème majeur pour le développement des applications qui peuvent être utilisées sur le Web.

2.3.3 La place de La langue arabe dans les systèmes de recherche d'information sur le web :

Les outils utilisés pour la recherche d'information en arabe sont peu nombreux puisqu'ils sont au nombre de deux :

- Recherche plein texte : la plupart des moteurs de recherche développés par des entreprises commerciales font de la recherche en intégral : parmi elles : www.alidrisi.com, le moteur de recherche de Sakhr ou encore www.alltheweb.com, moteur de recherche multilingue d'Unicode ou encore www.google.fr qui présente des versions multilingues dont l'arabe (El-Hachani, 2005).

- Recherche et indexation par analyse morphologique : cette option a pu se faire grâce au concours de la recherche académique, car elle demande des moyens humains et financiers importants pour construire des ressources linguistiques nécessaires à l'analyse linguistique de la langue arabe et pour le traitement des textes. On trouve notamment des analyseurs morphosyntaxiques, des bases lexicales et des dictionnaires (Abdelali et al., 2004 ; El-Hachani, 2005).

Il y a plusieurs projets de recherche qui ont pour objectif de modéliser des SRI en arabe citant en titre d'exemple d'après l'article de (Zaidi et al., 2005):

1. L'un de ces systèmes le MicroAirs system par Al Kharashi en 1991. Le travail d'Al Kharashi's était une investigation de trois principales méthodes de recherche nommées : le stem, racine ou mot en vue d'identifier le plus approprié

à l'arabe. Son étude indique une supériorité des méthodes de recherche par stem par rapport aux méthodes de recherche par mot dans l'exécution du rappel mais la précision est meilleure avec la méthode de mot (Zaidi et al., 2005).

2. AIRSMA (Arabic Information Retrieval System based on Morphological Analysis) est un autre étude par Al Tayyar dans sa thèse de PhD. Al Tayyar compare quatre : Racine, Stem, mot et la méthode morpho-sémantique, qu'il a développé. Les résultats de cette recherche sont de suggérer que les deux méthodes de recherche par racine et stem donnent une performance mieux que celles des méthodes du mot et la méthode morpho-sémantique en terme de précision (Zaidi et al., 2005).

2.3.4 Les ontologies et la langue arabe:

Il y a peu d'études publiées employant les technologies du web sémantique dans le développement des applications en langue arabe. Les recherches récentes explorant la potentialité des applications WS pour l'Arabe sont principalement dans le domaine Cross-Language Information Retrieval (CRIL) (Al-Khalifa et Al-Wabil, 2007).

Dans les travaux d'El-Helw et Aly en (2004) (El-Helw et Aly, 2004), une base de données d'applications intelligentes est développée par l'utilisation des techniques et structures de Web Sémantique. L'intégration du système développé dans cette étude est d'extraire les donnés à partir des documents structurés et non structurés disponibles dans les journaux en arabe. Ils ont élaborés des définitions formelles aux règles dans les domaines spécifiés pour permettre l'interrogation de ces données pour les informations implicites. L'étude est loin en ce sens, que seule une technologie clef de la SW a été employée dans le système intégré, à savoir les règles de l'inférence. SW contenant les principales technologies telles que les ontologies, pour représenter les données ont ajouté une dimension intéressante à leur approche (Al-Khalifa et Al-Wabil, 2007).

Zaidi et Laskri (2005) ont développé un outil web multilingue pour la recherche d'information en arabe basé sur une ontologie dans le domaine ligistatif. Ils ont utilise Protégé pour le développement de l'ontologie et ils ont aussi construit un moteur de requêtes pour la recherche d'information avec Protégé (Zaidi et al., 2005).

Une autre des utilisations du web sémantique pour la langue arabe est une expérience menée par Abdelali et al. (2003), dans laquelle les chercheurs ont recherche les documents en version arabe à partir d'un corpus multilingue. Ils ont utilisé les ontologies au lieu de l'habituelle technique lexicale de recherche d'information pour récupérer les documents en version arabe. Malgré les résultats prometteurs que l'ontologie fondée sur la recherche d'information technique a fourni, toute la puissance des technologies de SW n'a pas été démontrée dans leur recherche (Al-Khalifa et Al-Wabil, 2007).

2.4 Indexation des documents sur le web et recherche d'informations:

L'indexation et la recherche d'information sont très fortement liées, pour ne pas dire conditionnées l'une par l'autre. En effet, à quoi cela sert-il d'indexer des textes si les informations et leurs emplacements, repérés ne sont pas réutilisés par un système de recherche.

Un système de recherche fait correspondre la représentation de l'information que s'en fait l'utilisateur et celle utilisée par le système pour caractériser cette même information (Bessou et al., 2007).

Le premier problème dans l'indexation fut de déterminer les éléments que l'on doit choisir comme index. Dans les premiers projets sur la RI, on s'interrogeait sur les questions suivantes :

- indexation manuelle ou automatique ?

- vocabulaire libre ou contrôlé?

- quels mots à ajouter dans la stopliste ?

- quelles méthodes de troncature ? (Jian-Yun, 2008)

Ces choix dépendent de plusieurs paramètres, à savoir la taille du corpus, le domaine et l'application (Zargayouna, 2005).

2.4.1 Une pratique très ancienne :

La problématique fondamentale de l'indexation n'est pas nouvelle : des concordances bibliques aux moteurs de recherche actuels, en passant par les catalogues-matières et les divers index.

Schématiquement, trois grandes étapes dans l'histoire de l'indexation :

- le repérage de l'information dans les textes des manuscrits : Moyen-Age
- le repérage des documents eux-mêmes : de la Renaissance (imprimerie) aux banques de données
- le retour au texte intégral aujourd'hui, avec les outils d'indexation du texte intégral : numérisation du texte et démultiplication des possibilités de repérage et d'analyse de l'information (Serres, 2003).

2.4.2 Définition et objectifs d'indexation :

2.4.2.1 Définition de l'indexation :

La phase d'indexation est une phase importante, elle permet de représenter un document pour le rendre exploitable et manipulable par une recherche ultérieure (Bessou et al., 2007). Si un document est mal indexé, il risque de ne plus être retrouvé et donc perdu. La norme AFNOR NF Z 47-102 1996, définit l'indexation de la manière suivante :

"L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide des représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse" (Zargayouna, 2005).

Une deuxième définition est donnée par l'Association Française de normalisation comme suit : « Le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération » (Bessou et al., 2007).

2.4.2.2 Objectif de l'indexation:

Le but de l'indexation est d'extraire le contenu des documents et de les représenter de manière à identifier le document. Le choix des unités de représentation (appelées aussi unités d'index ou descripteurs) du document est crucial, il influence la qualité de l'indexation. L'ensemble de ces unités constitue le vocabulaire d'indexation, ce vocabulaire peut être libre ou contrôlé. Ces unités varient d'une chaîne de caractères (n-gramme) à des groupes nominaux ou des unités linguistiques complexes. La structure de l'index est reliée au choix du modèle de recherche utilisé (Zargayouna, 2005).

2.4.3 Types d'indexation

Il existe trois types d'indexation :

2.4.3.1 L'indexation manuelle :

Chaque document est analysé par un spécialiste du domaine ou alors par un documentaliste (Baziz, 2005), elle assure une bonne correspondance entre les documents et les termes. Cependant, cette méthode demande un travail manuel qui est non seulement très difficile mais très long à réaliser par les indexeurs (Abrouk, 2006).

Salton en 1986 a démontré les inconvénients de ce type d'indexation, par exemple : deux indexeurs peuvent indexer deux documents identiques avec des termes différents et des différences d'indexation peuvent également exister chez la même personne qui indexe a des moments différents.

L'indexation manuelle a l'avantage d'assurer une meilleure correspondance entre les documents et les termes choisis par les indexeurs pour les représenter (termes d'indexation). Ceci a pour conséquence une meilleure précision dans les documents que le système de RI retourne en réponses aux requêtes des utilisateurs. L'inconvénient majeur de cette méthode d'indexation est l'effort intellectuel qu'elle exige (en temps et en nombres de personnes). De plus, un degré de subjectivité lié au facteur humain fait que pour un même document, des termes différents peuvent être sélectionnés par des indexeurs différents. Il peut même arriver qu'une personne, à des moments différents, indexe différemment le même document (Baziz, 2005).

2.4.3.2 L'indexation automatique :

C'est lorsque le processus d'indexation est complètement informatise. Elle est réalisée en plusieurs étapes (Abrouk, 2006):

- l'extraction automatique des mots clés qui correspondent au mieux au contenu informationnel du document,
- l'élimination des mots vides ou mots fonctionnels (ex : conjonctions de coordination),
- la lemmatisation pour retrouver la racine des mots,
- la pondération des mots, pour affecter un poids élevé aux mots les plus importants.

L'indexation automatique est sans doute celle qui a été le plus étudiée en recherche d'information, étant donnée sa faculté d'automatisation du processus d'indexation. Elle comprend un ensemble de traitements sur les documents. On y distingue : l'extraction automatique des descripteurs, l'utilisation d'un anti-dictionnaire pour éliminer les mots outils, la lemmatisation, le repérage de groupes de mots, la pondération des mots avant de créer l'index. Ces traitements, mis à part le dernier (la pondération), sont détaillés dans le chapitre sur l'Indexation Conceptuelle (Baziz, 2005).

2.4.3.3 L'indexation semi-automatique :

Lorsqu'un premier processus automatique permet d'extraire les termes du document. Cependant le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations entre les mots clés et choisir les termes significatifs (Baziz, 2005). Les systèmes les plus simples et les plus répandus sont bases sur la sélection de mots-clés dans les textes (Abrouk, 2006).

Dans le cas de l'indexation semi-automatique (Jacquemin et al., 2002), appelée aussi indexation supervisée, les indexeurs utilisent un vocabulaire contrôlé sous forme de thesaurus ou de base terminologique. C'est le cas notamment lorsqu'il s'agit d'indexer des articles du domaine médical à l'aide du thesaurus MeSH. Les termes dans les bases terminologiques, sont préordonnés. Ils peuvent être liés par de simples relations

hiérarchiques tels que Is-a (pour former des taxonomies ou arbres) ou par un ensemble plus riche de relations lexicaux-sémantiques (dans ce cas on parle de réseau sémantique).

2.4.4 Langages documentaires :

Le terme langage d'indexation compte parmi ses synonymes : langage documentaire, langage contrôlé, etc. Et ce terme recouvre également de nombreux équivalents anglais : « indexing languags », « documentary language », « information retrieval language », etc (Sidhom, 2002).

Le langage d'indexation est un langage artificiel, c'est-à-dire construit à l'aide d'un ensemble de règles données, servant à la représentation abrégée du contenu d'un document. Dès lors, l'indexation consiste à détecter les termes les plus représentatifs du contenu du document (Hernandez, 2005).

On peut distinguer deux types de langage d'indexation:

- Langage contrôlé : il s'agit d'un lexique figé de descripteurs. L'indexation est alors le plus souvent manuelle, parfois semi-automatique, un professionnel choisit un ou plusieurs descripteurs pour représenter le document (Baziz, 2005).

Elle sert à éviter les problèmes d'ambiguïté (dus à l'homonymie et à la polysémie de certains termes) ainsi que les problèmes de redondance (synonymie, etc.). Ainsi, un terme d'indexation possède un seul sens défini et chaque sens donné n'est décrit que par un seul terme. Les descripteurs retenus seront les seuls mots clés acceptés lors de la requête. Une liste de termes interdits est aussi explicitement définie, ces termes ne doivent pas faire partie des descripteurs. Des relations peuvent être définies entre les termes retenus et d'autres termes (les non descripteurs). Le vocabulaire est généralement organisé dans un thésaurus où des relations entre descripteurs peuvent être représentées (e.g. relations d'équivalence, d'association, etc.) et permettent une représentation riche du document (Zargaouna, 2005).

L'indexation en langage contrôlé réduit le nombre de représentations possibles d'un document. Cela n'empêche pas l'indexation d'être subjective si elle est réalisée par un sujet humain, même si les sens et les termes sont bien délimités, l'interprétation humaine du contenu textuel est sujette à variation.

- Langage libre : est dit libre parce qu'il n'est pas contraint par un contrôle. Les descripteurs sont choisis librement et aucune contrainte n'est fixée a priori (Zargaouna, 2005). Les descripteurs sont extraits automatiquement des documents, ou de la requête de l'utilisateur. un document est le plus souvent indexé par la liste des mots qui le composent (Baziz, 2005). La représentation des documents est donc plus souple et permet une couverture large du contenu. Cette approche est néanmoins sujette aux problèmes d'ambiguïté sémantique de la langue naturelle. Le langage libre pose aussi le problème d'adéquation avec la requête, les mots clés recherchés doivent figurer parmi les mots clés choisis pour décrire le document. Le vocabulaire étant ouvert et non prédéfini, comment un utilisateur peut-il formuler une requête sans connaître les mots clés qui ont servi à indexer le texte ? Si l'indexation est manuelle, le risque de variabilité des descriptions est d'autant plus fort (Zargaouna, 2005).

2.4.5 Indexation des documents sur le web :

Cette étape primordiale doit s'effectuer avant l'étape de recherche effective de l'information, elle consiste à analyser le document lors de l'organisation du fond documentaire afin de produire un ensemble de mots clés, appelés aussi descripteurs, que le système pourra gérer aisément puis utiliser dans le processus de recherche ultérieur. Cette opération est appelée indexation.

Cet ensemble de mots clés peut être regroupé dans un thésaurus, mais en pratique, un thesaurus représente une notion plus large qu'une liste de mots clés. Il regroupe plusieurs relations de types linguistique (équivalence, association, hiérarchie) et statistique (pondération) (Baziz, 2005).

2.4.5.1 Extraction des termes :

La phase de sélection des termes d'indexation est fondamentale dans le processus de RI: ce sont en effet ces termes qui vont représenter les documents et requêtes au sein du SRI. Il convient donc de choisir ceux qui reflètent le mieux leur contenu sémantique.

Cette sélection est généralement liée à une phase de pondération décrite en section suivante. Dans l'idéal, les termes retenus doivent, d'une part, être le plus univoque et discriminant possible et, d'autre part, être en nombre limité à n de ne pas complexer les calculs effectués lors de la comparaison des représentations. Plusieurs traitements complémentaires peuvent être utilisés par les SRI pour pouvoir répondre à ces deux exigences (Moreau, 2006).

2.4.5.2 Pondération des termes :

La pondération est l'une des fonctions fondamentales en RI. Elle est la clé de voûte de la majorité des modèles et approches de RI proposés depuis les années 1960 (Baziz, 2005). L'objectif est de trouver les termes qui caractérisent le mieux le contenu d'un document. La manière la plus simple pour calculer le poids d'un terme est de calculer sa fréquence d'apparition, un terme qui apparaît souvent dans un document peut bien caractériser son contenu. Cependant, "l'informativité" d'un terme, c'est-à-dire, l'information sur le document que véhicule le terme, dépend aussi de sa répartition dans toute la base, on définit ainsi le pouvoir discriminatoire d'un terme, c'est-à-dire, le degré avec lequel un terme distingue un document d'un autre (Zargayouna, 2005).

La grande majorité des approches et systèmes opérationnels, se base sur les aspects statistiques. Ces méthodes tirent leur origine de la loi de Zipf et de la conjecture de Luhn.

2.4.5.2.1 Loi de Zipf :

La loi de Zipf est une loi empirique énoncée en 1949 par G.K Zipf. Elle décrit la répartition statistique des fréquences d'apparition des différents éléments d'un ensemble.

Selon Zipf, les mots dans les documents ne s'organisent pas de manière aléatoire mais suivant une loi inversement proportionnelle à leur rang. Le rang d'un mot est sa position dans la liste décroissante des fréquences des mots du corpus (Baziz, 2005). Formellement, cette loi s'exprime de la manière suivante :

$$\text{Rang} * \text{Fréquence} \approx \text{Constante}$$

La loi de Zipf a été par la suite étendue à d'autres domaines telles que la répartition des éléments (pixels) dans les images, des populations dans les villes ou encore, cas récemment, les pages web sur Internet, où

$$\text{Popularité d'une page} * \text{nombre d'accès à une page par mois} \approx \text{constante}$$

Dans le domaine de la recherche d'information, la loi de Zipf est utilisée pour déterminer les mots qui représentent au mieux le contenu d'un document. Pour cela, un autre concept est introduit, il s'agit de la conjecture de Luhn (Baziz, 2005).

2.4.5.2.2 La conjecture de Luhn :

Cette conjecture met en jeu le facteur d'informativité, autrement dit l'information contenue dans les termes. La juxtaposition des courbes de la fréquence et de l'informativité illustre cette conjecture (Luhn, 1958).

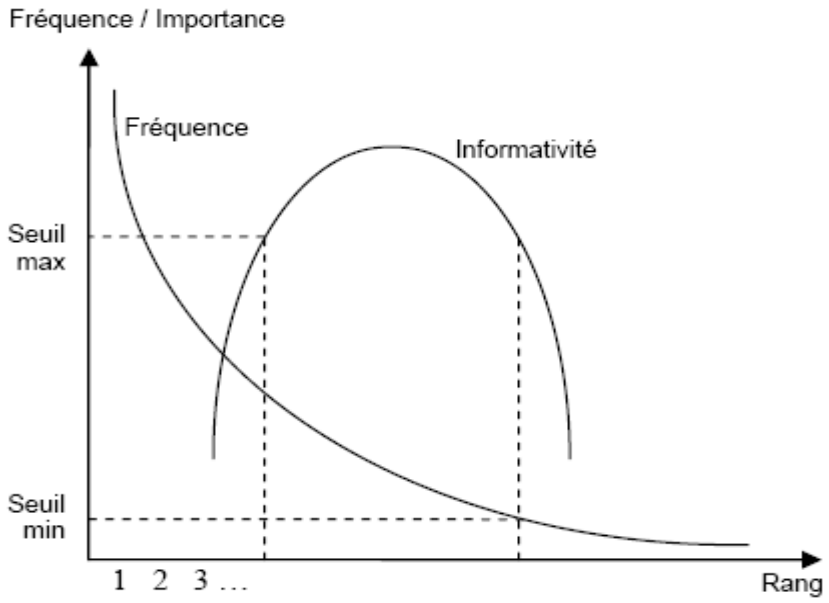

Figure 11. : La conjecture de Luhn.

Cette conjecture est utilisée pour diminuer la taille des index des documents. Deux seuils de fréquence sont fixés pour éliminer les termes dont le contenu informatif est jugé faible. Seuls les termes entre ces deux seuils sont alors pertinents pour représenter les documents.

2.4.5.2.3 Pondération en TF*IDF :

L'idée derrière la pondération des termes d'indexation est d'affecter aux termes d'un document, un poids pour traduire son importance dans le document, donc son degré d'informativité (Baziz, 2005).

La majorité des méthodes de pondération sont construites par la combinaison de deux facteurs. Un facteur de pondération local, et un second facteur de pondération globale.

Pondération locale :

Détermine l'importance d'un terme dans un document. Elle est, généralement, représentée par sa fréquence (tf) (Zargaouna, 2005).

TF signifie (Term Frequency), pour chaque terme t_j on calcule sa fréquence tf_{ij} dans le document d_i.

$$tf_{ij} = f(t_j,d)/Max[f(t,d_i)]$$

Où $f(t_j, d_i)$ représente la fréquence d'occurrence t_j dans le document d_i et $Max[f(t,d_i)]$ est la fréquence maximale des termes dans le document d_i.

Pondération globale

Mesurant la représentativité globale du terme vis-à-vis de la collection des documents.

Elle indique la représentativité globale du terme dans l'ensemble des documents de la collection.

IDF signifie (Inverted Document Frequency), il se calcule comme suit :

$$IDF = Log\ N/n$$

Où N est le nombre de documents dans le corpus, n est le nombre des documents qui contiennent le terme en question.

Plusieurs formules sont proposées pour le calcul du tf et du idf. Nous en présentons quelques unes. Dans le tableau suivant avec leurs images (intervalle des valeurs) (Zargaouna, 2005):

Les différentes fonctions tf		**Les différentes fonctions idf**	
Formule	**image**	**Formule**	**image**
$f(d,t)$	$[0, \quad +\infty]$	$\frac{1}{df(t)}$	$\left[\frac{1}{\lvert D\rvert}, \quad 1\right]$
$\frac{f(d,t)}{\max_{t`} f(d,t`)}$	$[0, \quad 1]$	$\log(1+\frac{\max_{d,t`} f(d,t`)}{df(t)})$	$[\log(2), \quad \log(1+cste)]$
$\frac{1}{2}+\frac{1}{2}\frac{f(d,t)}{\max_{t`} f(d,t`)}$	$\left[\frac{1}{2}, \quad 1\right]$	$\log(1+\frac{\lvert D\rvert}{df(t)})$	$[\log(2), \quad \log(\lvert D\rvert+1)]$
$1+\log(f(d,t))$	$[1, \quad +\infty]$	$\log(\frac{\lvert D\rvert}{df(t)})$	$[0, \quad \log(\lvert D\rvert)]$

Tableau 4. : Les différentes fonctions tf et idf.

Les données de base de ces formules sont f(d; t) qui est la fréquence du terme dans le document et df(t) qui est le nombre de documents ayant au moins une occurrence du terme t. Les fonctions tf dénotent une monotonie croissante et df une monotonie décroissante.

Du fait de cette double pondération (locale et globale) (Baziz, 2005), ces fonctions de pondération sont souvent référencées sous le nom de TF*IDF.

La mesure TF*IDF est une bonne approximation de l'importance d'un terme dans un document, particulièrement dans des corpus de documents de tailles homogènes. Les meilleurs termes étant ceux qui apparaissent fréquemment dans certains documents mais rarement dans le reste de la collection, un document pourra alors être représenté par ses seuls termes ayant un tf-idf suffisant (Boughanem, 2006).

2.4.5.3 Appariement documents- requête :

La requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur.

Le processus d'appariement document-requête permet de mesurer la pertinence d'un document vis-à-vis d'une requête (Baziz, 2005). De manière générale, à chaque réception d'une requête, le système crée une représentation similaire à celle des documents, puis calcule un score de correspondance entre la représentation de chaque document et celle de la requête. Ce score traduit un degré de pertinence système. Cette dernière est supposée représenter le jugement de pertinence de l'utilisateur vis-à-vis du document. La valeur de pertinence système est calculée à partir d'une fonction de similarité appelée (probabilité) RSV(Q, d) (Retrieval Status Value) où Q est une requête et D un document de la base. Cette mesure tient en compte des poids des termes déterminés généralement en fonction d'analyses statistiques et probabilistes (Baziz, 2005).

Le processus d'appariement est étroitement lié au processus d'indexation et de pondération des termes des requêtes et des documents du corpus. Notons également que d'une façon générale, le modèle de représentation des documents et requêtes ainsi que l'appariement document-requête, permettent de caractériser et d'identifier un modèle de recherche d'information.

Les principaux travaux effectués dans ce domaine ont fait l'objet de modèles de recherches d'information, citons ici les modèles de base (Hernandez, 2005):

- Le modèle booléen est basé sur l'algèbre de Boole et repose sur une représentation booléenne des requêtes. Dans ce modèle, les documents restitués à l'utilisateur sont ceux contenant exactement les termes de la requête. Il repose donc sur l'absence ou la présence des termes retenus pour indexer les documents et les termes de la requête.

- Le modèle vectoriel présenté par Salton repose sur les fondements mathématiques des espaces vectoriels. Dans ce modèle, les documents et les requêtes sont représentés sous forme de vecteurs dans l'espace des termes, issus de l'indexation. Les documents sont ensuite ordonnés à partir de leur ressemblance à la requête. Plusieurs mesures (Produit scalaire, Mesure de Dice, Mesure de Jaccard, …) permettent de calculer la similarité entre ces deux éléments correspondant aux calculs de la distance entre les deux vecteurs.

- Le modèle probabiliste repose sur la probabilité de pertinence d'un document connaissant la requête. Le modèle de langage mesure la probabilité de générer une requête à partir du modèle de langage du document.

Dans l'ensemble de ces modèles, les documents sont restitués à l'utilisateur par ordre de pertinence supposée décroissante.

2.4.6 L'indexation des documents guidée par les ontologies:

Parce que l'information sur le Web n'est pas facile a retrouver, compte tenu de l'hétérogénéité et la mise a jour de cette information (Abrouk, 2006), des approches basée sur l'indexation des documents Web avec des mots clés rattachés à des concepts d'ontologie sont conçus.

Parmi ces travaux prenant par exemple :

2.4.6.1 Indexation dans OntoSeek :

Le système OntoSeek de Guarino (Guarino et al., 1999) est l'une des premières tentatives expérimentales (1996-1999) sur l'application des ontologies à la RI, il est considéré comme une excellente référence dans ce domaine. Il est conçu pour traiter les pages jaunes et les catalogues de produits (en ligne) (Baziz, 2005). Le contenu des pages

ainsi que les requêtes sont modélisées par un formalisme basique de graphes conceptuels. OntoSeek combine un appariement basé sur le contenu et guidé par ontologie (ontology-driven content-matching) (Guarino et al., 1999).

Les objectifs de ce système sont (Guarino et al., 1999):

– L'utilisation des termes "arbitraires"du langage naturel (ontologie linguistique généraliste) qui permettent de d´écrire le contenu des documents ;

– Une flexibilité terminologique pour formuler les requêtes, grâce à un mécanisme d'intersection sémantique entre les requêtes et la description des produits ;

– Une assistance à la formulation, la généralisation ou la spécialisation des requêtes ;

– Des résultats précis et justes, et une efficacité raisonnable avec des volumes de données importants ;

– Une grande portabilité et extensibilité.

Différentes techniques de recherche d'informations ont été testées dans OntoSeek, avec :

- une liste de mots,
- une liste structurée de mots
- une liste de sens de mots avec une ontologie linguistique (Wordnet) et
- une liste structurée de sens de mots avec une ontologie linguistique (Wordnet).

Par ces techniques, OntoSeek a montré pour les pages jaunes l'intérêt d'utiliser une ontologie linguistique couplée avec une description du contenu structurée pour l'amélioration de la recherche d'informations (Abrouk, 2006).

2.4.6.2 Travaux de khan pour la désambiguïsation par les ontologies :

Khan (Khan, 2000), utilise la notion de concept, démarre de l'hypothèse qu'un groupe de mots-clés qui occurrent ensemble dans un même contexte détermine des concepts

appropriés pour désigner ensemble un autre concept, même si chaque mot-clé peut être individuellement ambigu.

En prolongeant et formalisant l'idée du contexte afin de réaliser la désambiguïsation des concepts, (Khan, 2000) propose un algorithme basé sur deux principes:

La co-occurrence et la proximité sémantique.

Cet algorithme de désambiguïsation, tache d'abord de désambiguïser à travers plusieurs régions de l'ontologie en utilisant le premier principe, et désambiguïse ensuite dans une région particulière en utilisant le second (Baziz, 2005).

On constate ici une couche supplémentaire de l'utilisation de l'ontologie, qui ne consiste pas juste a utiliser une ontologie pour seulement se restreindre a un domaine mais utiliser les relations inter-concepts afin d'exploiter les différents sens d'un terme dans un texte et supprimer les ambiguïtés (Abrouk, 2006).

2.4.6.3 Travaux de Desmontils et Jaquin pour l'indexation des pages web par les ontologies :

Le processus d'indexation semi-automatique des documents s'appuie sur des techniques issues du traitement automatique des langues et de l'ingénierie des connaissances (Desmontils et al., 2001). Dans le cadre de ces recherches, une suite d'outils a été développée qui permet d'effectuer une indexation structurée d'un site Web selon une ontologie donnée. Ces recherches ont mené à la détermination des processus suivants (figure 12) :

(i) Un processus qui extrait un ensemble de concepts candidats issus de pages Web, appelé index à plat.

Il permet d'extraire les termes bien formés issus des pages d'un site à l'aide d'analyses linguistiques (à l'aide de patrons morpho-syntaxiques) et de calculer un coefficient relatif à leur importance dans la page (celui-ci est fonction de la fréquence des termes et du poids des marqueurs HTML associés). Et d'autre part, il permet aussi de construire les concepts candidats représentatifs du contenu des pages à

l'aide du thesaurus WordNet et d'une mesure de similarité sémantique (qui prend en compte le contexte des concepts dans les pages).

(ii) Un processus concernant la désambiguïsation des labels des concepts de l'ontologie. Les pages Web sont des documents faiblement structurés et écrits en langage naturel. Pour faire le lien entre ces documents et les ontologies, un processus de désambiguïsation des labels associés aux concepts des ontologies a été déterminé. Ce processus s'appuie sur des heuristiques exploitant les relations de généralisation et de spécialisation présentes dans l'ontologie et les relations d'hyperonymie et d'hyponymie présentes dans une ontologie linguistique spécialisée construite à partir d'une ontologie issue du projet SHOE et du thesaurus WordNet (ils ont travaillé sur les universités américaines).

(iii) Un processus d'appariement des concepts candidats issus des documents et des concepts de l'ontologie.

À partir de l'ontologie linguistique, certains concepts candidats de l'index à plat sont retenus et les documents correspondants (les pages Web dans notre cadre d'étude) sont associés aux concepts correspondant de l'ontologie. Ceci permet de construire un index structuré des documents. La structure est donnée par l'ontologie. De nombreuses expérimentations et la mise au point de mesures d'évaluation spécifiques ont été effectuées.

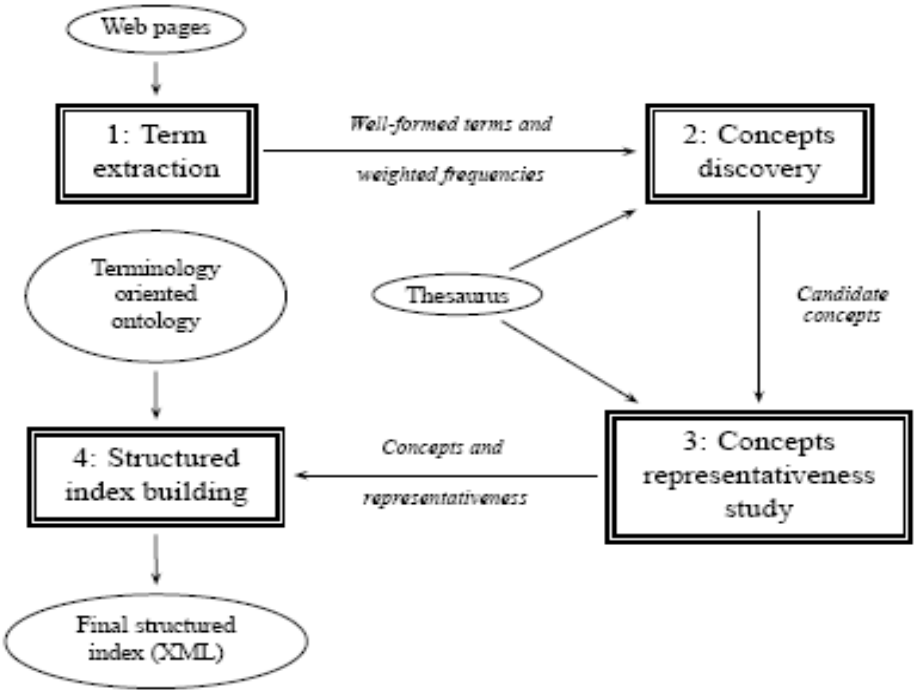

Figure 12. : Le processus d'indexation.

2.5 Conclusion :

Nous avons vu dans la première partie de ce chapitre les principales notions et concepts de la recherche d'information. Nous avons développé les principales étapes d'un processus de recherche d'information, Nous avons ensuite fait un panorama des outils de la recherche d'information. Par la suite, nous avons décrit l'utilité de l'utilisation des ontologies dans le domaine de RI, enfin on a évoqué quelques-unes des grandes tendances qui peuvent résumer les principaux bouleversements de la recherche d'information. Nous avons ensuite dressé un état de l'art sur l'évolution de la place de la langue arabe dans le domaine de la recherche d'information sur le web, passant par la définition de ses caractéristiques ainsi que le problème de son encodage et les solutions apportés. Après on a présenté les nouvelles recherches apporteront l'utilisation des ontologies avec la langue arabe.

Enfin on a fait un tour d'horizon sur l'indexation telle qu'elle se définit en recherche d'information. Débutant par la définition et les objectifs, passant par les types et les langages d'indexation. Ensuite on a aborde la section indexation des document sur le web, ou on a parle d'analyse de document par la description de l'extraction des termes

ou descripteurs, et la pondération ou l'objectif est de trouver les termes qui caractérisent le mieux le contenu d'un document. A la fin de cette section on a discuter les approches basée sur l'indexation des documents Web avec des mots clés liés a des concepts d'ontologie, ainsi que quelques exemples.

Dans le chapitre suivant, on va entamer et détailler la démarche de construction et utilisation des ontologie dans le domaine d'indexation des pages web arabes, avec quelques implémentations.

Chapitre 3 : Indexation d'un site web arabe avec une ontologie orientée terminologie.

3.1 Introduction:

Après avoir détaillé l'état de l'art des axes principaux de notre travail. Au sein de ce chapitre, on va présenter l'approche choisie pour l'utilisation des ontologies pour l'indexation des pages web arabes.

Dans ce contexte, on va mentionner la motivation du choix de l'approche. Après, les étapes de la construction de l'ontologie de domaine pour une finalité d'indexation ainsi que son implémentation qui va être établie. Ensuite une description détaillée de l'architecture est proposée pour l'outil d'indexation et des idées d'implémentation seront mentionnées. Et finalement on clôturant le chapitre par une conclusion.

3.2 Choix et motivation de l'approche de l'indexation des pages web arabes:

3.2.1 L'approche d'indexation choisie :

Les grands axes de l'approche sont inspirés des travaux de Desmontils et Jaquin, exposés dans (Desmontils et Jaquin, 2001), avec une projection sur la langue arabe pour atteindre les objectifs tracés dans notre mémoire. Les étapes de l'approche sont structurées sous forme d'un outil d'indexation basée sur les ontologies pour les pages web en arabe.

L'ontologie de domaine utilisée pour l'indexation des pages web est choisie à priori selon le contenu du site web. Dans notre travail, on va choisir un domaine spécifique pour illustrer l'approche, et on va ensuite construire une ontologie selon ce contexte.

Le processus est expliqué dans la section 2.4.6 du chapitre 2 et qui est illustré dans la figure suivante:

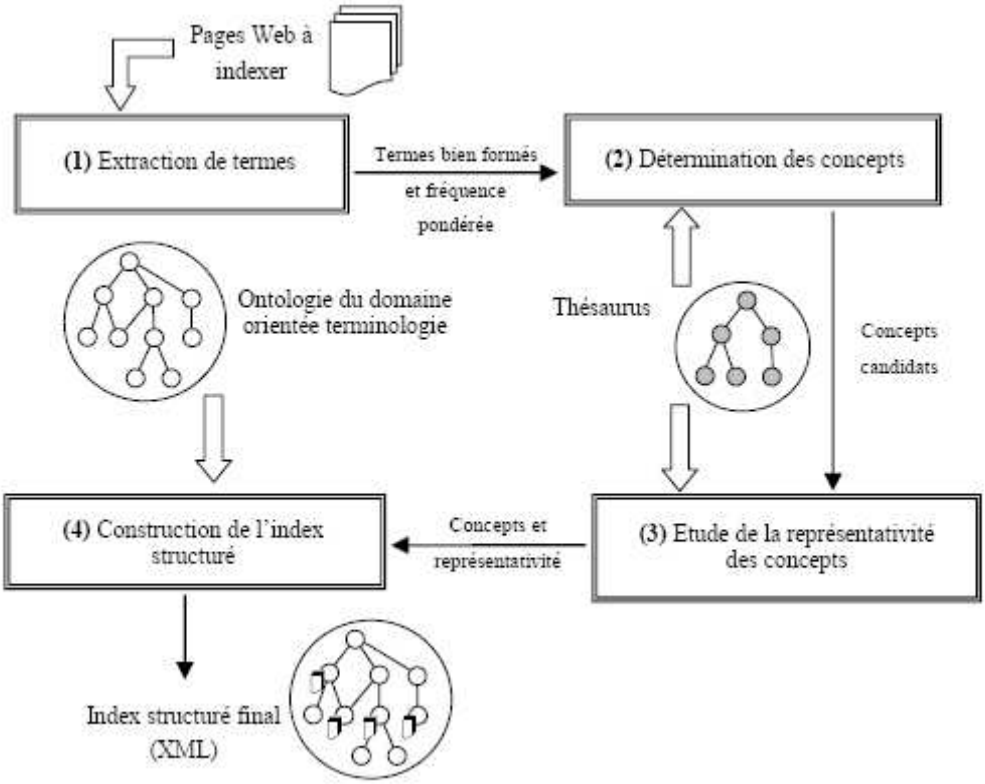

Figure 13. : Le processus générale d'indexation de Desmontils et Jaquin.

3.2.2 Motivation du choix :

Le processus d'indexation de Desmontils et Jaquin comprend un certain nombre d'avantages par rapport aux méthodes traditionnelles d'indexation, (seulement basées sur la recherche par mot-clé) et même sur les méthodes d'annotation de site Web, mentionnés dans (Desmontils et al., 2003):

- Certaines pages contiennent non seulement les mots clés mais aussi les concepts;
- Ces concepts sont représentatifs des sujets traités dans les pages sélectionnées;
- Les termes qui sont chargés de la page de sélection ne sont pas toujours ceux de la demande mais peuvent être synonymes;

- Les pages peuvent comprendre non seulement les concepts mais aussi d'autres plus spécifiques;
- L'importance d'un concept dépend non seulement de sa fréquence, mais également sur les marqueurs du code HTML qui le décrit et sur ses relations avec les autres concepts de la page ...

Le processus d'indexation peut être utilisé non seulement pour la recherche d'information mais aussi pour évaluer la pertinence d'un site Web à l'égard d'un domaine ou une connaissance.

En plus ce processus ne change pas les données du site web (page html, archive news ...) mais il crée un fichier XML additionnel (Desmontils et Jacquin, 2000).

3.3 Construction de l'ontologie de domaine أنطولوجيا_جامعة:

3.3.1 Choix d'une méthodologie de construction de l'ontologie:

Une méthodologie étant considérée comme ensemble de principes de construction systématiquement reliés, appliqués avec succès par un auteur dans la construction d'ontologies (Psyché, 2007). Alors nous avons tenté de construire une ontologie arabe dans le domaine des universités en fonction de mesures proposé par Noy et McGuiness dans (Noy et McGuinness, 2001), nous l'avons nommé أنطولوجيا_جامعة.

On va alors suivre les étapes suivantes, inspirées du guide de (Noy et McGuinness, 2001) avec des ajouts, des fusions et des éliminations des étapes vu inutiles:

Etape1: Définition du domaine et objectifs de l'ontologie : bien mentionner le domaine auquel l'ontologie va assister, plus les objectifs que l'ontologie va atteindre après son opérationnalisation.

Etape2: Définition des classes, de ces propriétés, et la hiérarchie des classes:

Il existe un certain nombre d'approches possibles pour développer une hiérarchie de classes:

- Un procédé de développement de haut en bas commence par une définition des concepts les plus généraux du domaine et se poursuit par la spécialisation des concepts.

- Un procédé de développement de bas en haut commence par la définition des classes les plus spécifiques, les feuilles d'une hiérarchie, et se poursuit avec le regroupement de ces classes en concepts plus généraux.

- Un procédé combiné de développement : est une combinaison des deux approches, de haut en bas et de bas en haut.

Remarquons que les types d'attributs: chaîne de caractère, entier, date doivent être définis à ce niveau.

Etape3: Définition des relations entre classes : définition des relations liant les différentes classes de l'ontologie (relation de spécialisation/généralisation, relation de classification…), ainsi que les cardinalités de ces relations (la cardinalité : définit le nombre de valeurs qu'un attribut peut avoir).

Etape4: Créer les instances: La dernière étape consiste à créer les instances des classes dans la hiérarchie. Ca exige de choisir une classe, créer une instance individuelle de cette classe, et la renseigner avec les valeurs des attributs.

Etape5 : Après l'accomplissement de ces étapes ; on va faire l'édition de l'ontologie avec un éditeur d'ontologie (Protégé, ontoedit…).

Toutes ces étapes vont être détaillées dans la suite.

3.3.2 Etapes de conception de l'ontologie **أنطولوجيا_جامعة** :

3.3.2.1 Définition du domaine et objectifs de l'ontologie :

Comme c'est mentionné auparavant, nous allons construire une ontologie dans le domaine des universités arabes. On va la nommer أنطولوجيا_جامعة.

3.3.2.1.1 Ressources d'acquisition des connaissances :

On s'est basé dans le processus d'acquisition des connaissances sur :

- Des interviews avec les experts de domaine, c'est-à-dire des administratifs et des enseignants.
- Consulter des sites web de quelques universités arabes, notamment les universités algériennes.
- Des inspirations des ontologies existantes dans le domaine, mais dans d'autres langues telle que l'ontologie de l'université de Washington.

3.3.2.1.2 Objectifs de l'ontologie أنطولوجيا_جامعة :

L'ontologie أنطولوجيا_جامعة est faite pour atteindre certains objectifs :

- Le premier objectif est la construction d'une l'ontologie أنطولوجيا_جامعة : une ontologie de concepts arabes, vu la rareté des ontologies arabes sur internet, et le besoin progressif de ces ontologies causé par leurs utilisations dans l'ingénierie cognitive en langue arabe.
- Le deuxième objectif de cette ontologie est de fournir une base de connaissances réutilisable à n'importe quelle université arabe, (notons que c'est un prototype).
- Le troisième objectif est que notre ontologie est construite pour illustrer notre hypothèse qui stipule que quelques principes du processus d'indexation de Desmontils et Jaquin sont valables aussi pour les pages en langue arabe. C'est la raison pour laquelle nous avons décidé de nous limiter à une ontologie de petite taille (quelques dizaines de concepts, relations et propriétés), même si nous sommes conscients qu'une ontologie est plus « puissante» quand elle est de grande taille.

3.3.2.2 Définition des classes et de leurs propriétés et la hiérarchie des classes:

3.3.2.2.1 Définition des classes et de leurs propriétés :

Vu le grand nombre des classes de notre ontologie, on va citer ci dessous quelques classes seulement, leurs propriétés et leurs métas classes :

La classe	Propriétés de la classe	Le méta classe
إصدار	إسم_الجامعة العنوان التاريخ التخصص المؤلف معلومات_إضافية	Le meta classe: الجامعة
تظاهرة علمية	إسم_الجامعة إسم_التظاهرة تاريخ_التظاهرة المدة المكان معلومات_إضافية	Le meta classe: الجامعة
شخص	إسم_الجامعة إسم_الشخص عنوان_الشخص رقم_الهاتف الحالة_العائلية معلومات_أخرى	Le meta classe: الجامعة
هيكل بيداغوجي	إسم_الجامعة الإسم تاريخ_الإفتتاح	Le meta classe: الجامعة
أطروحة	إسم_الجامعة الرقم_التسلسلي المشرف لجنة_المناقشة العنوان التاريخ التخصص	Le meta classe: إصدار

	المؤلف معلومات_إضافية	
أكاديمي	إسم_الجامعة التخصص_الأكاديمي إسم_الشخص عنوان_الشخص رقم_الهاتف الحالة_العائلية الدرجة الرتبة السلم الراتب تاريخ_التعيين معلومات_أخرى	Le meta classe: موظف
طالب	إسم_الجامعة رقم_التسجيل إسم_الشخص عنوان_الشخص رقم_الهاتف الحالة_العائلية التخصص_الدراسي معلومات_أخرى	Le meta classe: شخص
تقرير_في_مجلة	إسم_الجامعة العنوان إسم_المجلة التاريخ التخصص المؤلف معلومات_إضافية	Le meta classe: إصدار
مخبر_بحث	إسم_الجامعة	Le meta classe:

	الإسم تاريخ_الإفتتاح الباحث_المسؤول عدد_الباحثين عدد_الإصدارات	هيكل_بيداغوجي
……	……..	……

Tableau 5. : Définition des classes et leurs propriétés.

3.3.2.2.2 Etablissement d'une hiérarchie des classes de l'ontologie:

Nous avons utilisé une stratégie de haut en bas (top-down) pour la construction de la hiérarchie des concepts commençant par (جامعة=université).

Les relations entre les classes de la hiérarchie (ou taxonomie) sont des relations **is-a** pour exprimer l'hyponymie. Soient A, B : des classes :

A is-a B si tout les instances de B sont aussi instances de A.

La figure suivante illustre la hiérarchie de l'ontologie :

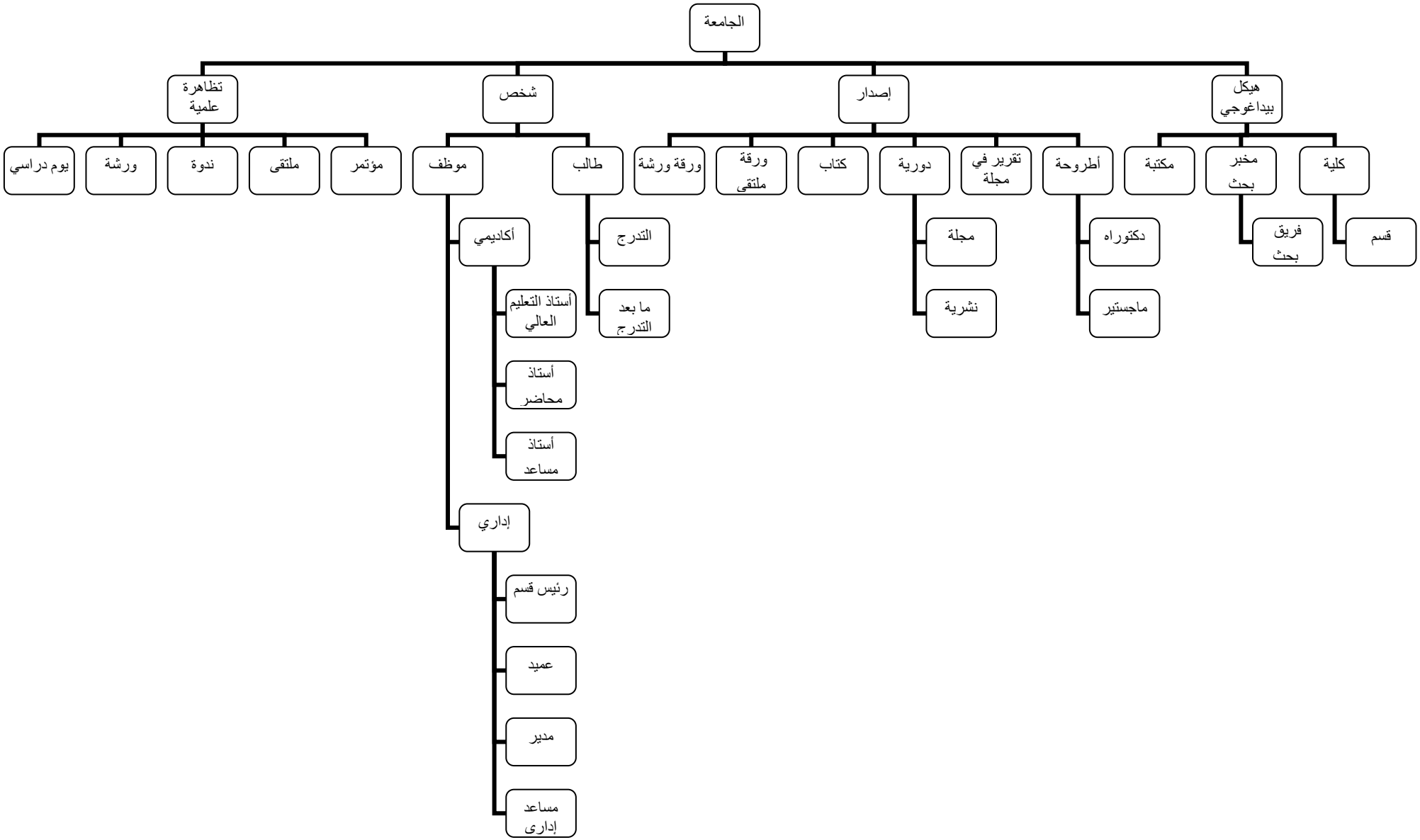

Figure 14. : La hiérarchie des concepts de l’ontologie.

3.3.2.3 Définition des relations entre les classes de l'ontologie:

On va récapituler cette phase par un diagramme UML (un diagramme de classes), qui est dans la littérature, une collection d'éléments de modélisation statique qui montre la structure d'un modèle.

Les différentes relations utilisées dans notre diagramme de classe sont:

- Classification.
- Association.
- Agrégation.
- Généralisation.
- Spécialisation.

Ils sont définit brièvement dans (Laallam, 2007).

Classification : les entités de même structure sont en général regroupées dans des classes, ce qui permet de manipuler l'ensemble des éléments possédant les mêmes caractéristiques.

Association : une association est un type de lien qui permet de relier plusieurs entités par un lien sémantique, en utilisant une association.

Agrégation : une agrégation est un type de lien qui permet de regrouper différentes entités en une nouvelle entité de niveau supérieur. L'identité de cette entité de niveau supérieur est déterminée par ses composants ou entités de niveau inférieur.

Généralisation : la généralisation est un type de lien qui consiste à créer une nouvelle entité (ou une nouvelle classe), à partir d'une union de plusieurs entités (ou classes) de niveau inférieur. Elle exprime un lien appelé IS.A entre un objet spécialisé et un objet générique.

Spécialisation : à l'inverse de la généralisation, la spécialisation est un type de lien qui consiste à créer une (ou plusieurs) sous-classe à partir d'une classe de niveau supérieur, en lui ajoutant des attributs propres.

En premier lieu et à titre d'illustration, le tableau suivant présente quelques relations entre les classes de notre ontologie:

Nom de la relation	Concept source	Cardinalité source	Concept cible	Cardinalité cible
يسير	مدير	(1,1)	جامعة	(1,1)
يشرف_على	أستاذ_محاضر	(1,1)	أطروحة	(0,n)
يحضر	ما_بعد_التدرج	(1,1)	ماجستير	(1,1)
ينظم	كلية	(0,n)	تظاهرة_علمية	(1,n)
Composition	مخبر_بحث	(4,n)	فريق_بحث	(1,1)
.....	.....	.....	.....	.....

Tableau 6. : Tableau de relations entre classes.

Remarque 1 : le diagramme de classes est édité a l'aide de l'outil Visual Paradigm for UML6.4

Remarque 2 : les propriétés des classes ne serons pas visualisées vu le grand nombre des classes et des relations.

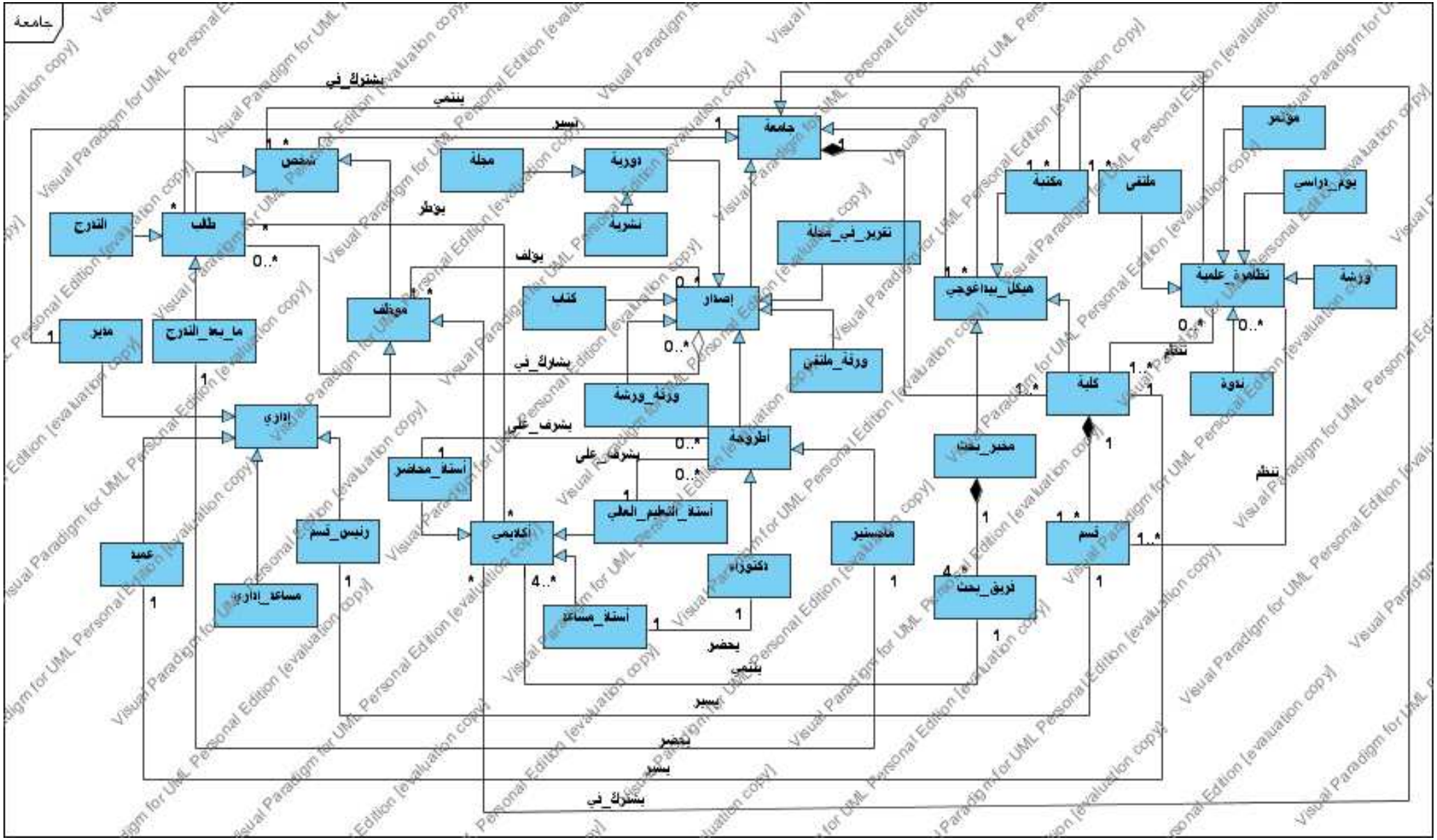

Figure 15Diagramme de classes.

Légende :

Représente une généralisation : est-un.

Représente une association porteuse de nom.

Représente une association de composition.

Représente une association d'agrégation.

3.3.2.4 Création des instances :

Ça vient, comme c'est déjà mentionné dans la description des étapes de construction de l'ontologie, à choisir une classe, créer une instance individuelle de cette classe, et la renseigner avec les valeurs des attributs.

A titre d'exemple, on va choisir deux instances, la première de la classe (أطروحة دكتوراه) et le deuxième de la classe (قسم).

Nom de l'instance	Attributs	Valeurs
نمذجة و تسييرصيانة الأنظمة الصناعية	إسم_الجامعة	جامعة باجي مختار عنابة
	الرقم_التسلسلي	؟؟؟
	المشرف	سلامي مختار
	لجنة_المناقشة	محمد الطيب العسكري
		زرهوني نورالدين
		شيخي سليم
		خلادي محمد خير الدين
		خدير طارق
	العنوان	نمذجة و تسييرصيانة الأنظمة الصناعية
	التاريخ	نوفمبر 2007
	التخصص	إعلام آلي
	المؤلف	لعلام فاطمة الزهراء
	معلومات_إضافية	لا شيئ
قسم الإلكترونيك	إسم_الجامعة	قاصدي مرباح ورقلة
	الإسم	قسم الإلكترونيك
	تاريخ_الإفتتاح	ديسمبر 2008

	عدد_الأساتذة عدد_الطلبة	25 210
……	……	……

Tableau 7. : Tableau des instances.

3.3.2.5 Etape5 : L'édition de l'ontologie :

L'édition de l'ontologie est une étape prémordiale pour passer à une ontologie opérationnelle spécifiée en utilisant un langage de représentation.

On a choisi dans le contexte de notre travail, l'éditeur d'ontologie Protégé 2000.

Protégé 2000 est un éditeur qui permet d'éditer une ontologie pour un domaine donné. Protégé 2000 est un éditeur de base de connaissance indépendant de la plate forme. C'est un environnement à base de frames pour le développement des systèmes à base de connaissances (Laallam, 2007). C'est l'outil le plus puissant qui supporte la langue arabe (Zaidi et al., 2005).

Une ontologie dans Protégé est structurée en classes, slots, facettes et axiomes. La base de connaissances de Protégé 2000 inclut l'ontologie et les instances individuelles des classes, avec des valeurs spécifiques des slots.

La figure ci-dessous montre notre hiérarchie de classes en utilisant Protégé 2000, après avoir saisir les classes et les attributs de l'ontologie. Elle est sous la forme d'un arbre :

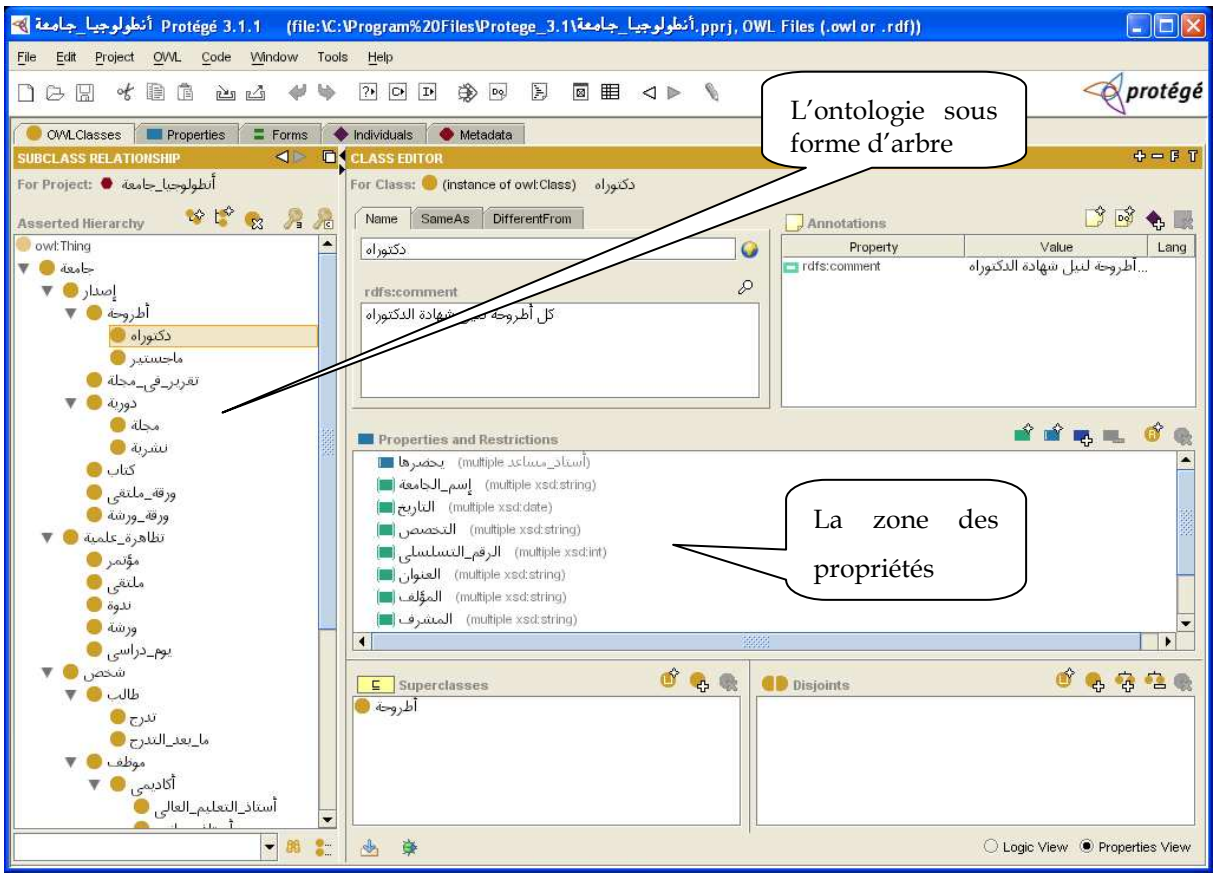

Figure 16. : Présentation de l'ontologie أنطولوجيا_جامعة dans Protégé2000.

Tandis que la figure ci-dessous montre l'ensemble des propriétés de l'ontologie أنطولوجيا_جامعة, qui sont saisies dans des formulaires fournis par Protégé 2000:

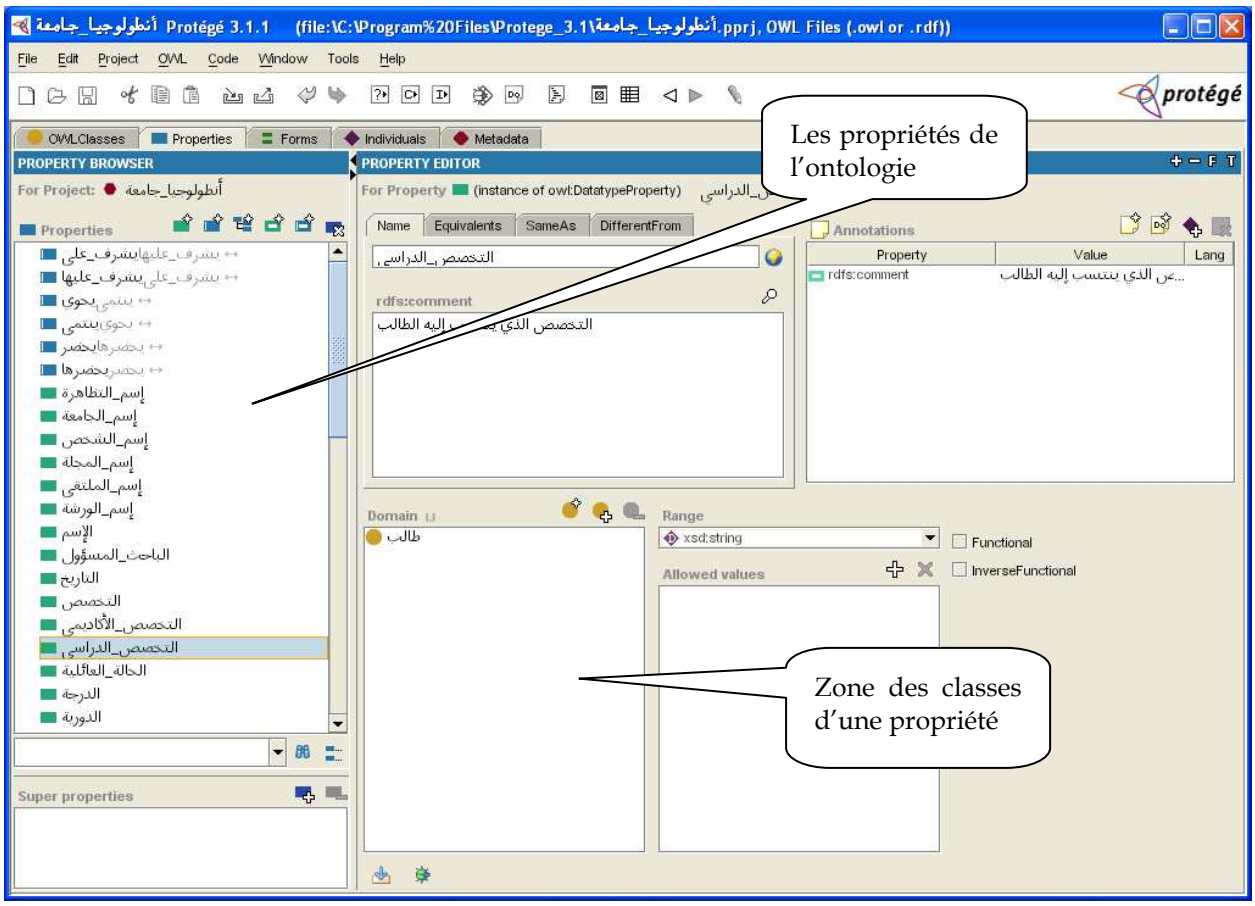

Figure 17. : L'ensemble des propriétés de l'ontologie أنطولوجيا_جامعة.

Une fois notre ontologie est construite et enregistrée, nous obtiendrons un document sous format OWL. Dans ce qui suit, nous présentons quelques déclarations.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns="http://www.owl-ontologies.com/أنطولوجيا_جامعة.owl#"
  xml:base="http://www.owl-ontologies.com/أنطولوجيا_جامعة.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="ورقة_ورشة">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ورقة_ملتقى">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_مساعد">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="أكاديمي"/>
    </rdfs:subClassOf>
```

```
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_التعليم_العالي">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#أكاديمي"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="تدرج">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="طالب"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="كتاب">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ماجستير">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="أطروحة"/>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل أطروحة لنيل شهادة الماجستتر</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_محاضر">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#أكاديمي"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="تقرير_في_مجلة">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >تقرير يقبل في مجلة معترف بها علميا</rdfs:comment>
.....
```

Figure 18. : Extrait du code OWL de l'ontologie أنطولوجيا_جامعة.

Dans l'**Annexe A** on trouvera un extrait de l'ontologie أنطولوجيا_جامعة.

3.3.3 Désambiguïsation des étiquettes des concepts de l'ontologie de domaine :

Afin de rendre possible la phase 4 mentionnée dans la démarche adoptée, nous allons mettre en place un processus concernant la désambiguïsation des étiquettes des concepts de l'ontologie. Ce processus s'appuie sur des heuristiques exploitant les relations de généralisation et de spécialisation présentes dans l'ontologie et les relations d'hyperonymie et d'hyponymie présentes dans le thésaurus. L'ontologie ainsi obtenue est appelée « *ontologie orientée terminologie* » du domaine.

Le thésaurus utilisé dans les travaux de Desmontils et Jaquin est WordNet (Miller, 1990). Il est vu comme une ontologie linguistique. Dans WordNet, un concept appelé un *sens,* est défini par un seul ensemble de *synonymes,* appelé *synset.* Les concepts dans le thésaurus sont ambiguës (il est associé à plusieurs termes), car pour chaque terme, le numéro propre du sens est donné après le symbole '#'.

Dans notre travail, on propose deux suggestions dans le but de désambiguïser les étiquettes l'ontologie arabe:

1. L'utilisation d'un thésaurus arabe, mais on a pas pu trouver un thésaurus complètement arabe. Néanmoins, El-Hachani dans (El-Hachani, 2005), a utilisé UNBIS thésaurus. Le multilingue UNBIS Thésaurus contient la terminologie utilisée dans le thème Analyse de documents et d'autres documents pertinents des activités et programme des Nations Unies.

2. L'utilisation du thésaurus WordNet accompagné d'un dictionnaire multilingue pour la traduction bidirectionnelle. Pour cet objectif, on propose le dictionnaire électronique multilingue incluant l'arabe : Tarjim de Ajeeb (disponible en ligne par Sakhr company software).

Nous avons opté pour la seconde suggestion, parce que nos propres essais des deux thésaurus ont révélé que WordNet est plus riche linguistiquement.

On peut accéder à WordNet par le biais de son API, JWNL (un Java api pour accéder au dictionnaire apparenté de WordNet).

3.4 Architecture et description détaillées de l'outil d'indexation des pages web arabes:

Cette section se focalise sur la description de notre outil d'indexation des pages web arabes, avec un détail sur ces composants et leurs fonctionnalités ainsi que quelques idées d'implémentation.

On a choisi à titre d'exemple un site web arabe de l'université d'El-Emir Abdelkader des sciences islamiques de Constantine (http://www.univ-emir.dz), parce que c'est un site totalement arabe (pour ne pas tomber sur le problème du multilinguisme).

3.4.1 Architecture globale de l'outil d'indexation :

Notre outil d'indexation des pages web arabes représente une partie d'un SRI arabe basé sur les ontologies pour l'indexation. Le SRI contient deux modules, celui d'indexation et l'autre de recherche d'information, il est mené d'une interface de saisie de requête et d'affichage des résultas après la tache de la recherche.

L'architecture de cet outil va être basée sur nos inspirations du processus d'indexation du Desmontils et Jaquin avec une adaptation compte tenu des pages web arabes. La figure suivante illustre l'architecture générale de notre outil d'indexation:

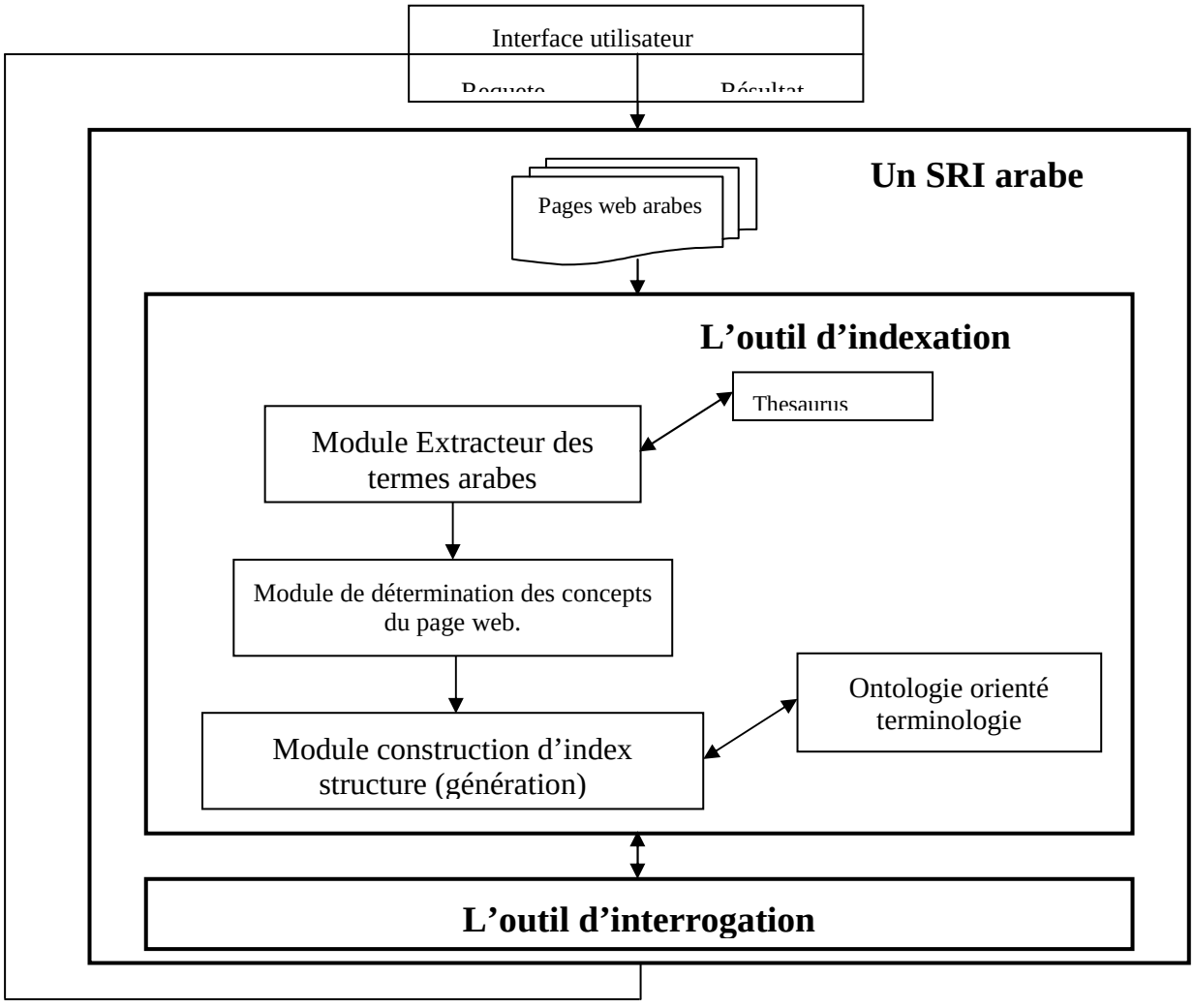

Figure 19. : Architecture générale de l'outil d'indexation (dans un SRI arabe).

L'outil d'indexation des pages web arabes se compose de 3 modules:

- Module d'extraction des termes arabes.
- Module de détermination des concepts.
- Module de construction d'index.

3.4.2 Description détaillée des modules de l'outil d'indexation :

3.4.2.1 Module1 : Extracteur de termes des pages web arabes.

C'est un ensemble de sous modules destinés à l'extraction automatique des termes des page web arabes, on se basant sur quelques techniques du traitement automatique de langue arabe.

Il a en entrée des pages web arabes qui vont être traitées automatiquement par le module d'extraction des termes. Après le processus, on aura comme résultat des termes bien formés pondérés. L'architecture détaillée du module d'extraction des termes arabes est présentée par la figure ci-dessous.

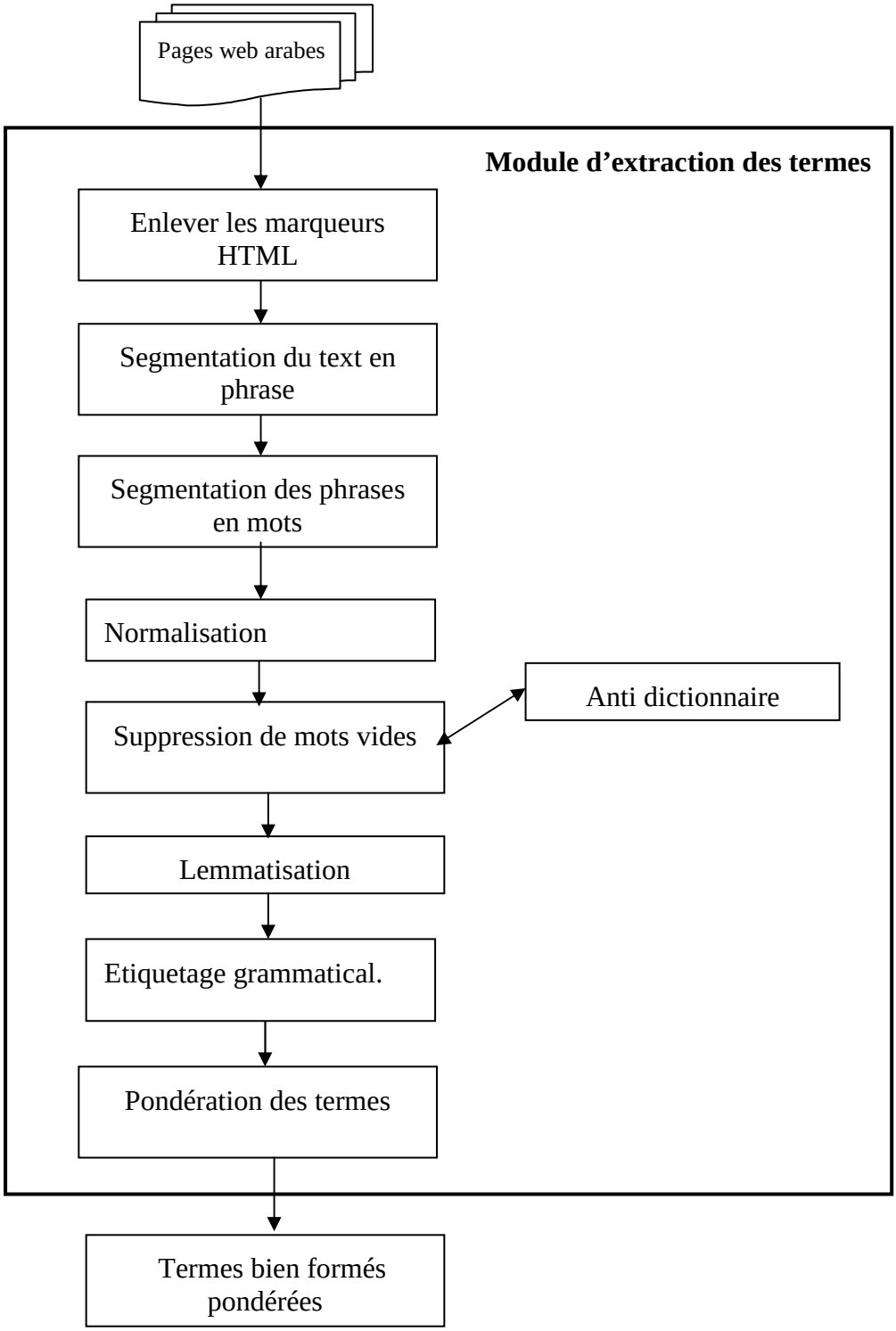

Figure 20. : architecture du module extraction des termes arabes.

Le processus l'extraction peut être décrite comme suit:

3.4.2.1.1 Enlever les marqueurs HTML :

Tous les marqueurs HTML des pages web seront enlevés. Alors on aura un texte brut (un ensemble de paragraphes). Ces marqueurs vont être utilisés ultérieurement pour le calcul des poids des termes.

3.4.2.1.2 Segmentation du texte en phrases :

Le problème de segmentation pour l'arabe réside dans la richesse grammaticale de cette langue (Baccour et al., 2003). Cependant, procéder à une analyse automatique d'un texte sans le segmenter en phrases peut conduire à des résultats non fiables. De même, avoir un mauvais segmenteur automatique de textes en phrases, conduit à accumuler les erreurs du traitement automatique du texte.

La reconnaissance de la fin de phrase est délicate car la ponctuation n'est pas systématique et parfois les particules délimitent les phrases (Douzidia, 2004).

Pour la segmentation de texte, il existe deux stratégies utilisées dans la littérature :

- Une segmentation morphologique basée sur la ponctuation,
- Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme : (أو= ou), (و= et), (أي= c.a.d), (لكن= mais), (حتى = quand).

Dans notre cas, l'identification des phrases au sein d'une page web se fait par la détection des balises <P> et </P>.

3.4.2.1.3 Segmentation des phrases en mots :

Appelée aussi tokenisation, elle consiste à séparer chaque phrase en une séquence de mots. Par la détection des délimiteurs de mots (tels que l'espace ou la ponctuation).

3.4.2.1.4 Normalisation :

La normalisation transforme une copie du document original dans un format standard plus facilement manipulable (Boulaknadel et al., 2008). Cette étape est considérée nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot arabe. Le document est normalisé comme suit :

- Suppression des caractères spéciaux ;
- Remplacement de إ, آ et أ avec ا;
- Remplacement de la lettre finale ي avec ى;
- Remplacement de la lettre finale ة avec ه.

3.4.2.1.5 Suppression des mots vides :

Consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec les éléments de l'anti-dictionnaire qui contient tous les mots non-significatifs appelés aussi des mots outils. Si un mot en fait parti, il ne sera pas pris en considération pour le calcul de sa fréquence. L'antidictionnaire regroupe en particulier les particules (comme : منذ, قبل,بعد, الذين …).

3.4.2.1.6 Lemmatisation :

La représentation suivante schématise une structure possible d'un mot (Douzidia, 2004). Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Post fixe	Suffixe	Corps schématique	Préfixe	Antéfixe

Pour détecter la racine d'un mot, il faut connaître le schème par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés.

Pour un mot significatif, on applique une lemmatisation légère qui consiste à essayer de déceler si des préfixes ou suffixes ont été ajoutés au mot (Darwish, 2002). Puisque la plupart des mots arabes ont une racine à trois ou quatre lettres, le fait de garder le mot au minimum à trois lettres va permettre de préserver l'intégrité du sens du mot.

La liste que nous utilisons regroupe les préfixes et les suffixes les plus utilisés dans la langue arabe, (Darwish, 2002) tels que les conjonctions, préfixes verbaux, pronoms possessifs, pronoms compléments du nom ou suffixes verbaux exprimant le pluriel etc.

On va utiliser la liste de préfixes et de suffixes proposé par (Darwish, 2003) montrée dans le tableau suivant :

Préfixes							
لا	في	للـ	كم	بم	وت	بت	وال
با	وا	ليـ	فم	لم	ست	يت	فال
	فا	ويـ	ال	وم	نت	مت	بال
Suffixes							
ا	ة	ين	ية	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	ية	نا	ها	كم	تي	ون

Tableau 8. : Liste des préfixes et suffixes les plus fréquents.

Par exemple, pour le mot arabe ايمان AymAn les préfixes possibles sont : "∅","A ا" et "Ay اي " et les suffixes possibles sont : "∅" et "An ان ", sans compter que ce mot peut aussi représenter un nom propre إيمَان Imène (Darwish, 2002).

Stem	Prefix	Template	Suffix	Root
ايمان "AymAn"	"#"	فيعال "CyCAC"	"#"	امن "Amn"
يمان "ymAn"	ا "A"	فعال "CCAC"	"#"	يمن "ymn"
مان "mAn"	اي "Ay"	فعل "CCC"	"#"	مان "mAn"
ايم "Aym"	"#"	فعل "CCC"	ان "An"	ايم "Aym"
يما "ymA"	ا "A"	فعل "CCC"	ن "n"	يما "ymA"

Tableau 9. : Les stems possibles pour le mot إيمَان.

3.4.2.1.7 Etiquetage grammatical :

Un étiqueteur et un sectionneur (tagger) qui parcours le flux de texte et assigne des étiquettes grammaticales de mots avec leur lemme (chaque mot aura sa catégorie grammaticale).

Desmontils et Jaquin (Desmontils et Jaquin, 2000) ont utilisé l'étiqueteur de Brill (Brill tagger). Pour notre travail, on utilise l'étiqueteur de Diab (Diab et al., 2004).

3.4.2.1.8 Pondération des termes :

Pour chaque terme bien formé résultants des étapes précédentes, est assigné un coefficient C, accordé à la fréquence du terme et son poids du marqueur HTML. Ce coefficient est appelé la fréquence pondérée (weighted frequency).

Par exemple, le marqueur "TITLE" a le poids10, "KEYWORD" a le poids 9.

Le tableau suivant contient quelques marqueurs HTML et leurs poids (Desmontils et Jaquin, 2000).

HTML marker description	HTML marker	Weight
Document title	<TITLE></TITLE>	10
Keyword	<meta name="keywords" ... content=...>	9
Hyper-link	<A HREF=...></A>	8
Font size 7	<FONT SIZE=7></FONT>	5
Font size +4	<FONT SIZE="+4"></FONT>	5
Font size 6	<FONT SIZE=6></FONT>	4
Font size +3	<FONT SIZE="+3"></FONT>	4
Font size +2	<FONT SIZE="+2"></FONT>	3
Font size 5	<FONT SIZE=5></FONT>	3
Heading level 1	<H1></H1>	3
Heading level 2	<H2></H2>	3
Image title	<IMG ... ALT="...">	2
Big marker	<BIG></BIG>	2
Underlined font	<U></U>	2
Italic font	<I></I>	2
Bold font	<B></B>	2
...	...	...

Tableau 10. : Quelques marqueurs HTML et leurs poids.

Dans une page Web contenant des termes différents m, pour un terme T, le coefficient *C (T)* est déterminé comme la somme pour les n occurrences du terme T de leurs poids de marqueurs HTML. Le résultat est normalisé. Ce calcul est indiqué dans la formule suivante, où $htmlcoefficient_i(T)$ correspond au poids marqueur HTML associés avec la ième occurrence du terme T.

$$c(T_k) = \frac{\sum_{i=1}^{n_j} htmlcoefficient_i(T_j)}{\max_{k\in[1,\ m]} \sum_{i=1}^{n_k} htmlcoefficient_i(T_k)}$$

Alors le résultat de ce module est : des termes bien-formés pondérés.

3.4.2.2 Module2 : Module de détermination des concepts du page web:

Nous nous disposons actuellement de termes bien formés et pondérés. Les termes bien-formés sont des formes qui représentent un concept particulier. Cependant, différentes formes peuvent représenter un même concept (par exemple : مديرالجامعة, رئيس الجامعة). Afin de déterminer non seulement le terme fixé inclus dans une page, mais aussi les concepts inclus dans une page, une ontologie linguistique est utilisée.

Une ontologie linguistique peut être considérée comme un dictionnaire structuré autour de groupes de mots synonymes, ce qui représente un concept. En outre, elle dispose de relations explicites entre les groupes de synonymes (relation hypernymie, relation meronymie...). Alors une ontologie linguistique est utilisée pour générer tous les concepts correspondant aux termes bien formés.

Dans notre expérience, nous avons utilisé le thésaurus WordNet (Miller, 1990) couplé avec le dictionnaire bilingue Tarjim de Ajeeb.

Le processus pour générer les synsets candidats est assez simple:

A partir de tous les termes extraits, tous les concepts candidats (tous les sens) sont générés en utilisant WordNet couplé avec le dictionnaire bilingue Tarjim. Ce thésaurus est d'une large couverture linguistique, mais il ne couvre pas tous les termes figurant dans la page web.

Si un terme n'existe pas dans WordNet, un sens spécifique est généré. Puis, le coefficient de convenance est calculé en utilisant une mesure de similarité sémantique. Il mesures pour un terme (une forme), la convenance possible avec tous les concepts qu'il pourrait représenter. Le calcul prend en compte les termes contexte (la page où il apparaît et ses voisins).

Il y a plusieurs mesures de similarité dans la littérature, celle de Rensink, de Wu & Palmer et d’autres.

Dans notre travail, on a adopté la mesure de similarité prise par Desmontils et Jaquin dans (Desmontils et Jaquin, 2000). Ceci pour des raisons liées à la spécificité des pages web.

Le coefficient de convenance , pour un concept spécifique, est la somme (sum) normalisée de tous les similarités sémantiques calculées avec tous les autres concepts inclus dans les pages Web étudiées.

Dans cette formule, un concept spécifique est unifié avec les synset correspondant dans WordNet. La mesure est indiquée dans les formules suivantes, où un terme T_k a l_k synsets associés, et il existe m termes dans les pages Web étudiés.

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

$$conv(synset_i(T_k)) = \frac{simsum(synset_i(T_k))}{\max_{j \in [1,\ l_k]} simsum(synset_j(T_k))}$$

3.4.2.3 Module3 : Module de construction d'index :

Nous nous disposons comme résultats des étapes précédentes, une ontologie avec les étiquettes de ses concepts disambiguisés, d'autre part les sens possibles dans chaque page HTML du site web arabe avec leurs fréquences relatives et leurs convenances évaluées.

Dans ce stade, les indexes sont liés avec les concepts de l'ontologie. Pour chaque sens, nous recherchons le même dans l'index. S'il existe, les coefficients sont ajoutés aux pages Web concernées.

3.5 Conclusion:

Ce chapitre était le dernier dans la démarche de notre travail, il a disposé d'un choix de démarche adoptée pour l'indexation sémantique des pages web arabes. Après, les étapes de construction de l'ontologie de domaine pour des finalités d'indexation sont établies. Par la suite, nous avons décrit l'architecture et la mise en place de notre outil d'indexation des pages web arabes, ainsi que ses modules (Module d'extraction des

termes arabes, module de détermination des concepts, et un module de construction d'index). Cet outil adopte la démarche choisie ultérieurement.

Notre démarche est semi-automatique. En effet, l'utilisateur a la possibilité d'intervenir d'une part sur la désambiguïsation des étiquettes dans l'ontologie et d'autre part sur le résultat de l'indexation.

Conclusion générale

Conclusion :

L'évolution très rapide d'Internet a permis l'émergence d'un savoir planétaire partagé mais a également généré plusieurs défis. En recherche d'information (RI), ceux-ci sont de trois types, à savoir, la gestion d'un volume d'informations, la présence de multiples supports et, finalement, le caractère multilingue de la toile. Dans ce dernier cas, l'importance grandissante d'autres langues que l'anglais a suscité le développement d'outils et de techniques automatiques afin de permettre leur traitement informatique. En comparaison à l'anglais et plus généralement aux langues sémitiques, la langue arabe présente des traits distinctifs, à savoir l'agglutination et la vocalisation.

Les travaux présentés dans ce mémoire se situent dans le contexte de la recherche d'information en langue arabe en utilisant les technologies du Web sémantique dans une perspective d'améliorer les performances des SRIs.

Ces travaux peuvent être présentés comme suit :

- Nous avons commencé par la construction d'une ontologie arabe orientée terminologie (أنطولوجيا_جامعة) pour être utilisée dans le processus d'indexation. Une opération de désambiguïsation est appliquée aux étiquettes des concepts de l'ontologie.
- Notre processus d'indexation est structuré sous forme d'un outil qui s'appuie sur des techniques de traitement automatique de la langue arabe pour l'extraction des termes bien formés depuis les pages web arabes. Ensuite les concepts associés aux termes sont générés. Les concepts d'index sont déduits par le biais d'une opération de mise en correspondance entre les concepts des pages web et ceux de notre ontologie orientée terminologie.

Perspective :

Les perspectives de notre projet porte principalement sur les points suivants :

1. Expérimentation et évaluation de l'outil d'indexation y compris ontologie arabe orientée terminologie pour mesurer ses impacts sur les performances des SRIs.
2. développement de la partie interrogation pour avoir un SRI arabe, vu le nombre restreint des SRIs Arabe basée sur les techniques du web sémantique. Et plus précisement l'utilisation de notre ontologie (أنطولوجيا_جامعة) dans la technique d'expansion de requêtes lors du processus d'interrogation de la RI.
3. Enrichissement de notre ontologie arabe (أنطولوجيا_جامعة) pour être multilingue.

Bibliographie

Abdelali A., Cowie J., Soliman H.S., (2004). Arabic Information Retrieval perspectives. JEP-TALN 2004, Arabic Language Processing - Text & Speech, Avril 2004. Disponible sur, http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/arabe2004/TAAA13.pdf

Abrouk L., (2006). Annotation de documents par le contexte de citation basée sur une ontologie. Thèse de doctorat, Université Montpellier II, soutenue 27 novembre 2006. Disponible sur, http://tel.archives-ouvertes.fr/docs/00/14/25/68/PDF/these.pdf

Al-Khalifa H., Al-Wabil A. (2007). The Arabic Language and the Semantic Web: Challenges and Opportunities. International Symposium on Computers and the Arabic Language, November 2007, Riyadh, Saudi Arabia. Disponible sur, http://hend-alkhalifa.com/wp-content/uploads/2008/02/29-en.pdf

Al-Khalifa H. S., Davis H. C., (2005). AraCore: An Arabic Learning Object Metadata for Indexing Learning Resources. MTSR, online, Spain, November 21-30, 2005. Disponible sur, http://eprints.ecs.soton.ac.uk/13220/01/AraCore.pdf

Anh T. T., (2005). Web sémantique et réseaux sociaux-construction d`une mémoire collective par recommandations mutuelles et (re-) présentations. Ecole nationale supérieure des télécommunications, juillet 2005.

Ashburner M., Ball CA., Blake JA., (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium, 2000.

Baccour L., Hadrich Belguith L., Mourad G., (2003). Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs. GEI'2003, Troisièmes Journées Scientifiques des Jeunes Chercheurs en Génie Electronique et Informatique, Ecole Nationale d'Ingénieurs de Sfax et l'Institut Supérieur d'Informatique et du Multimédia de Sfax, Mahdia, Tunisie, 18-20 mars. Disponible sur, http://www.regim.org/publications/conferences/2003/2003_GEI_LLG.pdf

Baeza-Yates R., Ribeiro-Neto B., (1999). Modern Information Retrieval. ACM Press, New York, 1999, chapter1 et chapter10. Disponible sur, http://www.dcc.ufmg.br/irbook/

Baget J. F., Canaud E., Euzenat J., Said-Hacid M., (2003). Les langages du web sémantique. mars 2003. Disponible sur, http://www.revue-i3.org/hors_serie/annee2004/revue_i3_hs2004_01_02.pdf

Bahloul D., (2006). Une approche hybride de gestion des connaissances basée sur les ontologies : application au incidents informatiques. Thèse de doctorat, L'institut national des sciences appliquées de Lyon, soutenue le 15 décembre 2006. Disponible sur, http://docinsa.insa-lyon.fr/these/2006/bahloul/these.pdf

Baziz M., (2005). Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, Université Paul Sabatier Toulouse, Soutenue le 14 décembre 2005.

Beckett D., (2003). RDF/XML Syntax Specification, (2003). Disponible sur, http://liris.cnrs.fr/alain.mille/enseignements/Ecole_Centrale/projets_2004/RDF.pdf

Benoit L., (2007). Notion d'ontologie et construction d'ontologie à partir de corpus de textes, Synthèse de lectures. Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal, février 2007. Disponible sur, http://www.benoit-lavoie.ca/public/docs/Notion_d_ontologie_et_ construction_d_ontologie_a_partir_de_corpus_de_textes.pdf

Berners-Lee T., Hendler J., Lassila O., (2001). Semantic Web. Scientific american, 284(5) :35-43, 17 mai 2001. Disponible sur, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

Bessou S., Saadi A., Touahria M., (2007). Un système d'indexation et de recherche des textes en arabe (SIRTA). Premier séminaire national sur le langage naturel et l'intelligence artificielle, LANIA'2007, Université Hassiba Ben Bouali, Chlef, 20-21 novembre 2007. Disponible sur, http://lania.site.voila.fr/Bessou.pdf

Biebow B., Szulman S., (1999). Terminae: A Linguistic-Based Tool to Build of a Domain Ontology. In Proceedings of the 11th European Knowledge Acquisition Workshop (EKAW'99), 1999. Disponible sur, http://gollem.swi.psy.uva.nl/workshops/ka2-99/camready/biebow.pdf

Bos B., (2002). XML en 10 points. W3C Communication Team, février 2002.

Boughanem M., (2006). Introduction à la Recherche d'Information. Laboratoire IRIT, Université Paul Sabatier de Toulouse, EARIA' 2006.

Boulaknadel S., Daille B., Aboutajdine D., (2008). Utilisation des termes complexes dans un système de recherche d'information en langue arabe. 9th International Conference on the Statistical Analysis of Textual Data, a Ecole normale supérieure Lettres et sciences humaines, Lyon, France, March 12 - 14 2008. Disponible sur, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/boulaknadel-daille-aboutajdibe.pdf

Bourdoncle F., (1999). Panorama et perspectives des outils de recherche d'information textuelle sur Internet. In : IDT 1999 : textes des communications. Disponible sur, http://www.exalead.com/Francois.Bourdoncle/idt99.html

Boutemedjet S., (2004). Web Sémantique et e-Learning. Cours IFT6261, Université de Monterial, 2004. Disponible sur, http://www.abhatoo.net.ma/index.php/fre/content/download/2202/22634/file/Web%20S%C3%A9mantique%20et%20e-Learning.pdf

Brickey D., Guha R., (1999). Ressource Description Framework Schema specification. W3C Communication Team, 1999. Disponible sur, http://www.w3.org/TR/PR-rdf-schema

Broder A., Henzinger M., (1998). Information retrieval on the Web: tools & algorithmic issues. FOCS 98 tutorial (1998). Disponible sur, http://www.research.digital.com/SRC/personal/Monika_Henzinger/focs-talk-m78/index.htm

Cardoner L., (2004). Rapport de Stage au niveau de l'IRIT. Institut de recherche en informatique de Toulouse, IUP, Université Paul Sabatier Toulouse III, 1er avril – 31 août 04. Disponible sur, http://foveaproject.free.fr/DESS.0903.pdf

Connolly D., Harmelen F. V., Horrocks I., McGuinness D. L., Patel-Schneider P. F., Stein L. A., (2001). DAML+OIL Reference Description. W3C Recommendation, December 2001. Disponible sur, http://www.w3.org/TR/daml+oil-reference

Darwish K., (2002). Building a Shallow Arabic Morphological Analyzer in One Day. Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA. Disponible sur http://www.aclweb.org/anthology-new/W/W02/W02-0506.pdf

Darwish K., (2003). Probabilistic Methods for Searching OCR-Degraded Arabic Text. Doctoral dissertation, University of Maryland.

Desmontils E., Jacquin C., (2000). Web Site Indexation and Ontologies. Rapport de recherche, université Nantes. Disponible sur, http://www.sciences.univ-nantes.fr/irin/Vie/RR/RR-IRIN-012.ps

Desmontils E., Jacquin C., (2001). Indexing a Web Site with a Terminology Oriented Ontology. International Semantic Web Working Symposium (SWWS'2001), Stanford, CA, USA, july 30 - august 1, 2001, pp. 549-565.

Desmontils E., Jacquin C., Morin E., (2002). Indexation sémantique de documents sur le web : application aux ressources humaines. In Proceedings of Journees de l'AS-CNRS Web semantique, Octobre 2002.

Desmontils E., Jacquin C., Simon L., (2003). Ontology enrichment and Indexing Process. RR-IRIN-03.05, Nantes, Mai 2003.

Diab M., Hacioglu K., Jurafsky D., (2004). Automatic tagging of arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004, Boston, pages 149-152.

Douzidia F. S., (2004). Résumé automatique de texte arabe. Mémoire présenté à la Faculté des études supérieures, en vue de l'obtention du grade de M.Sc en informatique, Université de Montréal, Septembre 2004. Disponible sur, http://rali.iro.umontreal.ca/Publications/files/DouzidiaMemoire.pdf

El-Hachani M., (2005). Indexation des documents multilingues d'actualité incluant l'arabe— équivalence interlangues et gestion des connaissances chez les indexeurs. Thèse de Doctorat, Université Lumière Lyon 2, France, Soutenue le 14 novembre 2005.

El-Helw A., Aly H., (2004). An Intelligent Database Application for the Semantic Web, in proc. CSITeA-04 conference.

Euzenat J., (2003). Qu'est-ce que le web sémantique!?. Inria, 2003.

Fensel D., Horrocks I., Broekstra J., Decker S., Erdmann M., Goble, Harmelen F. V., Klein M., Staab S., Studer R., Motta E., (2000). OIL: The Ontology Inference Layer. Technical Report IR-479, Vrije University Amsterdam, Sept. 2000.

Furst F., (2004). Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation. Thèse de doctorat, École des Mines de Nantes, soutenue le 25 Novembre 2004.

Gaelle L., (2002). État de l'art Ontologies et Intégration/Fusion d'ontologies. Une partie d'un rapport de stage, laboratoire Dialogue et Intermédiations Intelligentes de la Direction des Interactions Humaines DIH/D2I au (FTR&D) à Lannion, septembre 2002.

Ghafour S. A., (2003). Méthodes et outils Pour l'intégration des ontologies. Mémoire de stage de DEA, laboratoire LIRIS Lyon, 2003-2004.

Gruber T. R., (1991). The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. In Proceedings of the Second International Conference Principles of Knowledge Representation and Reasoning, (KR & R-91), 1991.

Gruber T. R., (1993). A translation approach to portable ontology specifications, Knowl. Acquis, 5(2) : 199–220, 1993.

Guarino N., (1997). Understanding, building and using ontologies. International J. Human-Computer Studies, pp 293-310. 1997.

Guarino N., (1998). Formal ontologies and information systems. In Guarino N., editor, Proceedings of FOIS'98, IOS Press, Amsterdam, 1998.

Guarino N., Masolo C., Vetere G., (1999). OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems, vol. 14, no. 3, pp. 70-80, May/Jun, 1999.

Heflin J., Hendler J., Luke S., (1999). Applying Ontology to theWeb : A Case Study. In International Work-Conference on Artificial and Natural Neural Networks (IWANN), 1999.

Heijst G. V., Schreiber A. Th., Wielinga B. J., (1997). Using explicit ontologies for KBS development. International Journal of Human-Computer Studies, 1997.

Hernandez N., (2005). Ontologies de domaine pour la modélisation du contexte en recherche d'information. Thèse de doctorat, Université de Paul Sabatier Toulouse, soutenue le mardi 6 décembre 2005. Disponible sur, http://www.irit.fr/~Nathalie.Hernandez/nHernandez.pdf

Hernandez N., Mothe J., (2007). An approach to evaluate existing ontologies for indexing a document corpus. Visité décembre 2007.

Jacquemin C., Daille B., J. Royauté J., Polanco X., (2002). In Vitro Evaluation of a Program for Machine-Aided Indexing. 2002.

Jian-Yun N., (2008). Le domaine de recherche d'information–Un survol d'une longue histoire. Département d'informatique et recherche opérationnelle, Université de Montréal. Consulte en février 2008. Disponible sur, http://www.iro.umontreal.ca/~nie/IFT6255/historique-RI.html

Khan L., (2000). Ontology-based Information Selection. Ph.D. thesis, University of Southern California, August 2000. Disponible sur, https://129.110.10.36/research/esc/publications/lkhan_def.pdf

Laallam F. Z., (2007). Modélisation et gestion de la maintenance dans les systèmes de production. Thèse de Doctorat, Université Badji Mokhtar Annaba, soutenue le Novembre 2007.

Laublet P., Reynaud C., Charlets J., (2002). Sur quelques aspects du Web sémantique. Aux assises Gdb 13 Nancy, 2002. Disponible sur, http://sis.univ-tln.fr/grd13/

Levini J., (2002)., the internet minute : languages on the net.

Luhn S., (1958). Automatic Creation of Literature Abstract. IBM Journal of Research and Development, 2(2): pp. 159-165, 1958. Disponible sur, http://www.research.ibm.com/journal/rd/022/luhn.pdf

Maedche A., Motik B., Silva N., Mafra R. V., (2002). A mapping framework for distributed ontologies. In Proceedings of the International Conference EKAW'2002, volume 2473, pages 235–250, Springer-Verlag LNAI, 2002.

McBride B., Packard H., (2004). RDF Primer. W3C Recommendation, 10 February 2004. Disponible sur, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

McGuinness D. L., Harmeln F. V., (2004). OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. Disponible sur, http://www.w3.org/TR/owl-features/

Mestiri A., (2007). Vers une approche web sémantique dans les applications de gestion de conférences. Mémoire de maîtrise en informatique, Université Laval, Québec, 2007.

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K., (1990). Introduction to WordNet: An On-line Lexical Database. (Revised August 1993). Princeton University, New Jersey. Disponible sur, http://wordnet.princeton.edu/5papers.pdf

Moreau F., (2006). Revisiter le couplage traitement automatique des langues et recherche d'information. Thèse de doctorat, Université de Rennes 1, soutenue le 07 décembre 2006. Disponible sur, http://www.irisa .fr/texmex/people/moreau/publications/these.pdf

Noy N. F., McGuinness D. L., (2001). Développement d'une ontologie 101 : Guide pour la création de votre première ontologie. Université de Stanford, Stanford, CA 94305, 2001 (traduction de l'anglais par Angjeli A., BnF, Bureau de normalisation documentaire.). Disponible sur, http://www.bnf.fr/PAGES/infopro/normes/pdf/no-DevOnto.pdf

Phan Q. T. T., (2005). Ontologies et web services. Rapport d'intérêt personnel encadré, Institut de la Francophonie pour l'Informatique, Hanoi, juillet 2005.

La Création d'Ontologies Web Sémantique avec Protégé-2000. Article disponible sur l'URL http://www.cetic.be/internal.php3?id_article=138, vu en novembre 2007.

Psyché V., (2007). Rôle des ontologies en ingénierie des EIAH : cas d'un système d'assistance au design pédagogique. Thèse Doctorat en informatique cognitive, Université du Québec à Montréal, Canada, soutenue juillet 2007. Disponible sur, http://hal.archives-ouvertes.fr/docs/00/19/00/48/PDF/Psyche-Valery-PhD-These-2007.pdf

Ranwez S. C., (2000). Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur. Thèse de doctorat, Université de Montpellier II, Soutenue le 21 décembre 2000.

Schuurman A., (2005). Recherche de services bioinformatiques dans une ontologie. Mémoire présenté en vue de l'obtention du grade de Maître en Informatique, Facultés Universitaires Notre-Dame de la Paix, 2005.

Serres A., (2003). Introduction a l`indexation, [en ligne]. Paris : 15 janvier 2002, Date de modification : 1 septembre 2003, CNDP, SavoirsCDI, Juin 2004. 14 p. Disponible sur, http://www.uhb.fr/urfist/Supports/Indexation/Indexation1Panorama.html#1.%20Panorama%20de%20l%E2%80%99indexation

Serres A., (2004). Recherche d'information sur Internet : où en sommes-nous, où allons-nous?. Paris : CNDP, SavoirsCDI, Juin 2004. 14 p. Disponible sur, http://savoirscdi.cndp.fr/culturepro/actualisation/Serres/Serres.htm

Sidhom S., (2002). Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherché d'information: de l'écrit vers la gestion des connaissances. Thèse de Doctorat, Université Claude Bernard Lyon1, France, soutenue le 11 mars 2002. Disponible sur, http://www.enssib.fr/bibliotheque-numerique/document-923

Sure Y., Angele J., Staab S., (2002). OntoEdit: Guiding Ontology Development by Methodology and Inferencing. In Proceedings of the Confederated International Conferences CoopIS, DOA and ODBASE, 2002.

Uschold M., King M., (1995). Towards a Methodology for Building Ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995), 1995.

Vignaux G., (2007). La recherche d'information : Panorama des questions et des recherches. CNRS-MSH Paris Nord, consulter décembre 2007. Disponible sur, http://plate-forme-ast.mshparisnord.org/IMG/pdf/La_recherche_d_info.pdf

Zaidi S., Laskri M.T., Bechkoum K., (2005). A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain. Proceedings IEEE International Conference on Signal-Image Technology and Internet-Based Systems, IEEE SITIS, 27 Nov–2 Dec, Yaounde, Cameroun pp 86-91. ISBN 2-9525435-0.

Zargayouna H., Salotti S., (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. IC 2004 : 15es journées francophones d'ingénierie des connaissances, Lyon France, 5-7 mai 2004. Disponible sur, http://liris.cnrs.fr/~ic04/programme/articles/Zargayouna-IC2004.pdf

Zargayouna H., (2005). Indexation sémantique de documents XML. Thèse de doctorat, Université Paris XI Orsay, Soutenue le 15 Décembre 2005.

Annexes

Annexe A : Extrait de l'ontologie أنطولوجيا_جامعة.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns="http://www.owl-ontologies.com/أنطولوجيا_جامعة .owl#"
  xml:base="http://www.owl-ontologies.com/أنطولوجيا_جامعة .owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="ورقة_ورشة">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ورقة_ملتقى">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_مساعد">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="أكاديمي"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_التعليم_العالي">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#أكاديمي"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="تدرج">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="طالب"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="كتاب">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ماجستير">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="أطروحة"/>
    </rdfs:subClassOf>
```

```
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل أطروحة لنيل شهادة الماجستتر</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="أستاذ_محاضر">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#أكاديمي"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="تقرير_في_مجلة">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >تقرير يقبل في مجلة معترف بها علميا</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="يوم_دراسي">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="تظاهرة_علمية"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="جامعة">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >هيكل بيداغوجي يشرف على العملية التعليمية لطلبة التدرج و ما بعد التدرج</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="موظف">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="شخص"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="عميد">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="إداري"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="مدير">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إداري"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="مساعد_إداري">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إداري"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="مؤتمر">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#تظاهرة_علمية"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="دورية">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إصدار"/>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >إصدار عن الجامعة بشكل دوري</rdfs:comment>
  </owl:Class>
```

```
  <owl:Class rdf:ID="فريق_بحث">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="مخبر_بحث"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="رئيس_قسم">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#إداري"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="هيكل_بيداغوجي">
    <rdfs:subClassOf rdf:resource="#جامعة"/>
  </owl:Class>
  <owl:Class rdf:ID="قسم">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="كلية"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="نشرية">
    <rdfs:subClassOf rdf:resource="#دورية"/>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >نشرية تصدر  عن الجامعة بشكل دوري</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="ندوة">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#تظاهرة_علمية"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#مخبر_بحث">
    <rdfs:subClassOf rdf:resource="#هيكل_بيداغوجي"/>
  </owl:Class>
  <owl:Class rdf:ID="ملتقى">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#تظاهرة_علمية"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#كلية">
    <rdfs:subClassOf rdf:resource="#هيكل_بيداغوجي"/>
  </owl:Class>
  <owl:Class rdf:ID="مكتبة">
    <rdfs:subClassOf rdf:resource="#هيكل_بيداغوجي"/>
  </owl:Class>
  <owl:Class rdf:ID="مجلة">
    <rdfs:subClassOf rdf:resource="#دورية"/>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >مجلة تصدر  عن الجامعة بشكل دوري</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="ما_بعد_التدرج">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#طالب"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#إداري">
    <rdfs:subClassOf rdf:resource="#موظف"/>
  </owl:Class>
  <owl:Class rdf:about="#إصدار">
    <rdfs:subClassOf rdf:resource="#جامعة"/>
```

```
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل ما يصدر عن الجامعة</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="دكتوراه">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#أطروحة"/>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل أطروحة لنيل شهادة الدكتوراه</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:about="#تظاهرة_علمية">
    <rdfs:subClassOf rdf:resource="#جامعة"/>
  </owl:Class>
  <owl:Class rdf:about="#طالب">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل شخص يتلقى تعليما عاليا    </rdfs:comment>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#شخص"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ورشة">
    <rdfs:subClassOf rdf:resource="#تظاهرة_علمية"/>
  </owl:Class>
  <owl:Class rdf:about="#شخص">
    <rdfs:subClassOf rdf:resource="#جامعة"/>
  </owl:Class>
  <owl:Class rdf:about="#أطروحة">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >كل إصدار يقدم أمام لجنة مناقشة لنيل شهادة عليا</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#إصدار"/>
  </owl:Class>
  <owl:Class rdf:about="#أكاديمي">
    <rdfs:subClassOf rdf:resource="#موظف"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="يحضرها">
    <rdfs:range rdf:resource="#أستاذ_مساعد"/>
    <owl:inverseOf>
      <owl:ObjectProperty rdf:ID="يحضر"/>
    </owl:inverseOf>
    <rdfs:domain rdf:resource="#دكتوراه"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="ينتمي">
    <owl:inverseOf>
      <owl:ObjectProperty rdf:ID="يحوي"/>
    </owl:inverseOf>
    <rdfs:range rdf:resource="#فريق_بحث"/>
    <rdfs:domain rdf:resource="#أكاديمي"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="يشرف_على">
    <rdfs:domain rdf:resource="#أكاديمي"/>
    <rdfs:range rdf:resource="#أطروحة"/>
    <owl:inverseOf>
      <owl:ObjectProperty rdf:ID="يشرف_عليها"/>
    </owl:inverseOf>
  </owl:ObjectProperty>
```

```
 <owl:ObjectProperty rdf:about="#يحضر">
   <rdfs:domain rdf:resource="#أستاذ_مساعد"/>
   <owl:inverseOf rdf:resource="#يحضرها"/>
   <rdfs:range rdf:resource="#دكتوراه"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:ID="يسير">
   <rdfs:range rdf:resource="#قسم"/>
   <rdfs:domain rdf:resource="#رئيس_قسم"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:about="#يشرف_عليها">
   <rdfs:domain rdf:resource="#أطروحة"/>
   <rdfs:range rdf:resource="#أكاديمي"/>
   <owl:inverseOf rdf:resource="#يشرف_على"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:ID="تحتوي_على">
   <rdfs:domain rdf:resource="#تظاهرة_علمية"/>
  <rdfs:range rdf:resource="#إصدار"/>
</owl:ObjectProperty>
 <owl:ObjectProperty rdf:about="#يحوي">
   <rdfs:range rdf:resource="#أكاديمي"/>
   <rdfs:domain rdf:resource="#فريق_بحث"/>
   <owl:inverseOf rdf:resource="#ينتمي"/>
 </owl:ObjectProperty>
 ...
```

Annexe B : WordNet

WordNet est un dictionnaire électronique de l'anglo-américain, développé depuis 1985 et initialement conçu pour tester les déficits lexicaux dans des expériences de psychologie cognitive.

Il a été développé à Princeton par George A. Miller, et continue à être mis à jour. Sa structure est celle d'un thésaurus, il est organisé autour de la structure des synsets, c'est-à-dire des ensembles de synonymes et de pointeurs décrivant des relations vers d'autres synsets. Chaque mot peut appartenir à un ou plusieurs synsets, et à une ou plusieurs catégorie du discours suivantes : nom, verbe, adjectif, adverbe.

Nous ne nous intéressons dans ce rapport qu'aux noms de WordNet, pour des raisons de simplicité. Notons que la base de connaissance de WordNet est composée de plusieurs graphes de concepts : en effet, il y plusieurs racines, ou concepts les plus généraux, et donc plusieurs graphes, censés se couper le moins possible. Les racines sont :

- entity ;
- location ;
- abstraction ;
- event ;
- group ;
- phenomenon ;
- psychological_feature ;
- shape ;
- state ;
- act ;
- possession.

Chacun des graphes dont nous venons de préciser la racine a un ensemble de concepts différent : de 43950 concepts pour entity a 688 pour shape.

WordNet est disponible gratuitement sur l'URL : http://wordnet.princeton.edu/