REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGN SUPERIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE 20 AOUT 1955-SKIKDA

FACULTE DES SCIENCE

DEPARTEMENT INFORMATIQUE

MEMOIRE

Pour l'obtention du diplôme de Master

Spécialité : Génie logiciel avancée et application

Thème

La recherche d'information dans le web basée sur ontologie

Présenté par :

- Abdelbaki Maroua
- Mokrane Imene

Dirigé par :

Dr.Benoudina Lazhar

Session 2021 - 2022

Remerciements

Nous remercions le bon Dieu de nous avoir mis sur la route du savoir

Nous remercions infiniment notre promotrice : Dr.Benoudina Lazhar, pour nous avoir encadrés durant cette année, Nous tenons à lui exprimer notre profonde gratitude.

Nous remercions également les membres du jury pour nous avoir honorées avec leurs présences et pour avoir jugé notre travail.

Nous tenons à saluer la peine et l'effort fournis par l'ensemble de nos professeurs afin d'assurer notre formation tout au long de notre cursus universitaire et leur disons de ce fait, merci. Nous exprimons notre infinie gratitude à nos chers parents en reconnaissance de leurs sacrifices, dévouement, soutien et encouragements.

Enfin, nous remercions tous ceux qui ont contribué de près ou de loin à réaliser ce modeste travail.

Ces quelques mots ne traduisent guère tout ce que nous avons pu recevoir de la part de chacun d'entre eux, mais nous souhaitons néanmoins qu'ils y trouvent l'expression de notre infinie reconnais

Dédicace

Je dédie cet événement marquant de ma vie

***A*mon père** *disparu trop tôt. J'espère que du mande qui est sien maintenant, il apprécie cet humble geste comme preuve de reconnaissance de la part d'une fille qui a toujours prié pour le salut de son âme .Puisse Dieu, le tout puissant, l'avoir en sa sainte miséricorde !*

A *ma maman* ***Attar Nacira*** *qui m'a soutenu et encouragé durant c'est années d'études. Qu'elle trouve ici le témoignage de ma profonde reconnaissance.*

A *mes très chers frères* ***Wassim*** *et* ***Assil*** *et mes belles sœurs* ***Abir*** *et* ***Rym.***

P*uisse Dieu vous donne santé, bonheur, courage et surtout réussite.*

Imene

Dédicace

Nous voilà à la fin d'un parcours si long mais plein de bons événements, de meilleurs souvenirs et d'inoubliables amitiés tissées tout au long de nos années d'études ;

Ce mémoire n'est pas seulement la finalité d'une année d'étude mais le résultat de tant d'années de travail et de recherche du savoir ;

Cet humble travail que nous allons présenter ci-après, n'a pu avoir lieu si tous ceux qui me sont chers ne se pas consacrés et venus à mon aide.

Je dédier mon modeste travail à :

M**a famille :* *Mon cher père* ***Mohamed*, grâce à lui j'ai réalisé mon rêve et ma chère mère* ***Djilani Akila****, qui mon jamais laissé sentir de manque et qui me ont pris soin de moi depuis ma naissance.*

M**on cher frère* ***Billel*, et mes belles sœurs* ***Rayane*** *et* ***Zina.***

M**es chers oncles et **M**es chères tantes :* ***Salah*,* ***Mourad****,* ***Hicham****,* ***Aicha,*** ***Fadila****,* ***Fatima****,* ***Mounira****,* ***Habiba****,* ***Nabila****,* ***Leïla****, et leurs enfants*

M**es belles amies :* ***Souad*,* ***Hadil****,* ***Mouna****,* ***Bouchra****,* ***Wided****, ……la liste est grande.*

C'est pour cela que je leur dédie mon travail en leur disant :

≪Je veux aimer et merci beaucoup pour tout ce que tout ce que vous m'avez offert≫

Maroua

Table des Matières

Liste des figures

Liste des tableaux

Introduction générale

Le Web sémantique (plus techniquement appelé « le Web de données ») permet aux machines de comprendre la sémantique, la signification de l'information sur le Web. Il étend le réseau des hyperliens entre des pages Web classiques par un réseau de lien entre données structurées permettant ainsi aux agents automatisés d'accéder plus intelligemment aux différentes sources de données contenues sur le Web et , de cette manière , d'effectuer des tâches (recherche , apprentissage , etc.) plus précises pour les utilisateurs . Le terme a été inventé par Tim Berners - Lee, Co - inventeur du Web et directeur du W3C, qui supervise l'élaboration des propositions de standards du Web sémantique. La plupart du temps, lorsque l'on prononce le terme de Web sémantique, on parle des différentes technologies qui se cachent derrière. Parmi les plus connues , on peut citer RDF (Ressource Description Framework) qui correspond à un modèle d'information , et les formats d'échanges de données en RDF pour communiquer entre différentes applications (RDF / XML , RDF / JSON , N3 , Turtle , N - Triples et d'autres) . Dans le domaine du Web sémantique , la sémantique des données est décrite par des ontologies avec des langages prévus pour fournir une description formelle de concepts , termes ou relations d'un domaine quelconque . Ces langages sont RDFS (Ressource Description Framework Schema) et OWL (Web Ontology Language). Il existe aussi des langages de description des données structurées dans du XHTML afin que des outils effectuent un traitement automatique de ces différentes données. Ces langages sont RDFa et Micro format et, nouvellement arrivé avec HTML 5, Microdata ... Ensuite, pour finir avec la liste des technologies, il existe un langage de requête, au même titre que SQL pour les bases de données relationnelles, SPARQL, qui effectue des requêtes mais sur des triplets RDF. Il en existe d'autres (RQL et RDQL), mais ils sont bien moins utilisés.

Problématique

Bien que l'information ne soit pas statique, et sujette à des modifications, des enrichissements, s'altère avec le temps, et parvient de différentes sources, nous avons besoin d'outils et de modèles qui permettent aux utilisateurs et experts du domaine de constituer, consulter et maintenir leurs connaissances du domaine. Ainsi, pour faire face à l'importante masse d'informations sur le web, il y a un besoin accru de développer des outils et des techniques qui servent à trouver les documents contenant l'information pertinente pour un besoin bien spécifique.

Un moyen de parvenir à améliorer la recherche d'informations sur le web est le marquage sémantique des ressources en utilisant des ontologies qui définissent les concepts du domaine et leurs relations.

Objectif de notre travail

. Notre travail est basé sur l'utilisation d'ontologie OWL pour la recherche des informations dans le web.

Domaines concernés :

Ce mémoire situe à la croisée de plusieurs domaines de recherche tels que l'Ingénierie des connaissances, l'Ingénierie ontologique, la recherche d'information, la modélisation informatique, l'informatique médicale, la gestion et l'optimisation des coûts en médecine.

Nous sommes intéressés au la recherche d'information sur le web et l'Ingénierie ontologique avec l'utilisation des techniques de cette dernière, dans le cadre de l'informatique médicale.

Nous espérons que le travail réalisé sera profitable, et nous présentons brièvement ces quatre principaux champs de recherche auxquels ce travail apporte sa contri1bution la plus significative.

Organisation du mémoire : Ce mémoire contient 4 chapitres :

Chapitre 1 : est consacré à la recherche d'information sur le web. Nous traitons trois points, à savoir, les éléments distinctifs entre la RI classique et la RI sur le web, les sources d'information spécifiques aux documents web et les différentes méthodes ou modèles développés pour combiner ces différentes sources d'information dans le but d'améliorer la pertinence de la recherche d'information.

Chapitre 2 : Qui s'intitule Ingénierie Ontologique, nous avons défini la notion d'ontologie ainsi que les composants de cette dernière, et nous avons examiné les différents types d'ontologies et les méthodes de construction, nous avons aussi consacré une partie dans ce chapitre sur les outils et langages d'ontologies et enfin quelques domaines dont les ontologies interviennent.

Chapitre 3 : présente la construction d'une ontologie de domaine partons de connaissances brutes et arrivant à une ontologie opérationnelle, ce processus est inspiré de la méthode ' METHONTOLOGY ' qui contient cinq étapes (´Evaluation des besoins, Conceptualisation, Formalisation, Implémentation et Maintenance). Cette ontologie permet aux utilisateurs et experts humaines (Médecins, étudiants, Patients...) de constituer, consulter et maintenir leurs connaissances sur ce domaine, et d'avoir un langage médical commun et standardisé.

Chapitre 4 : Ce dernier s'intitule Implémentation et réalisation, nous avons détaillé encore plus sur le logiciel utilisé (Protégé 4.3), et ensuite nous présenterons notre ontologie Implémentée sous l'éditeur protégé qui sera interrogée par des requêtes afin de tester la consistance de cette dernière et expliqué les différentes étapes de la mise œuvre de notre site.

Chapitre 01

La recherche d'information sur le web

1.1 Introduction

Les problématiques posées en recherche d'information sur le web sont identiques à celles posées par la recherche d'information classique (indexation, appariement, etc.). Dès, l'apparition du web, différentes catégories d'outils se sont développées pour répondre à ces problématiques et pour faire face aux nouveaux chalenges posés par le web. La spécificité du web réside dans le type de documents manipulés (pages HTML), qui sont des documents structurés, le nature hypertexte du web et le nombre d'utilisateur et le type de requête qu'ils utilisent pour exprimer leur besoin en information. Généralement, ces outils de RI sur le web examinent la combinaison de diverses sources d'information telles que : le contenu textuel du document et la structure du web, pour classer les documents web en réponse à une requête utilisateur. Ce chapitre présente un aperçu sur la RI sur le web. Dans la première section, nous présentons les différences entre la RI classique et la RI sur le web. Dans la seconde section, nous examinons les différentes sources d'information spécifiques au document web : la structure interne des documents web et la structure des liens. Nous décrivons dans la troisième section les différentes approches utilisées pour combiner les sources d'information sur un document web, dans le but d'améliorer la pertinence de la recherche d'information.

1.2 Différences entre la RI classique et la RI sur le web

Les SRI classiques sont souvent développés et utilisés dans des environnements bien contrôlés tel que les bibliothèques, où les collections de documents sont généralement des petites tailles et les utilisateurs ont des besoins en informations bien spécifiques.

Le web diffère sur plusieurs points avec les autres ressources documentaires rencontrées habituellement en recherche d'information. Parmi les facteurs distinctifs, on peut citer, le volume du web; la dispersion, l'hétérogénéité et la nature dynamique de l'information dans cet espace ; enfin, les utilisateurs du web proviennent de divers horizons avec des niveaux de connaissances différents et expriment leur besoins en information avec peu de mots.

Nous analysons brièvement ci-dessous ces facteurs distinctifs.

1.2.1 Le volume du web : Le volume d'informations accessible sur le web ne se mesure plus en giga-octets mais en tira octets voir en péta-octets et hexa-octets. Déjà, à la fin de l'année 1995, le moteur Altavista avait reporté qu'il a indexé approximativement 30 millions de pages web statiques. En juin 2000, Netcraft[5] recense plus de 12 millions de sites web, ce qui

représente environ 800 millions de pages web. En Mai 2013, une étude a rapporté que le moteur de recherche Google comporte dans son index plus de 46 milliards de pages web[6].

D'un autre côté, chaque mois, plus de 100 milliards de recherches sont effectuées, au travers des moteurs de recherche commerciaux sur le web **[1]**.

Cette augmentation de la taille du web et le nombre de requêtes soumises sont à l'origine de la dégradation des performances des processus de recherche tant en terme d'efficience que d'efficacité. Plus précisément **[2]:**

- L'allongement des délais de réponse;
- L'augmentation des temps d'indexation;
- La diminution de la précision de la recherche.

1.2.2 L'hétérogénéité de l'information

L'hétérogénéité des ressources d'information sur le web inclut plusieurs points : les ressources du web sont écrites dans plusieurs langues (une centaine). Une variante de formats sont utilisées, rien que pour le texte on peut citer : HTML, PDF, XML, RTF, DOC, etc., et elles utilisent différents encodages dans la plupart du temps ils sont incompatibles.

L'hétérogénéité peut être aussi sémantique, tous les thèmes sont traités sur le web, et ces thèmes sont abordés par diverses sources, qui peuvent être des sources scientifiques, de vulgarisations ou de commercialisations. Cette hétérogénéité créée de nouveaux défis significatifs pour la recherche d'information qui sont : l'interopérabilité entre sources d'information et l'amplification des phénomènes de polysomie et d'homographie, qui ont pour effet d'augmenter le bruit lors d'une recherche.

1.2.3 La disparité de l'information

La disparité est une caractéristique qui traduit l'occurrence disséminée de l'information dans de larges collections de documents. Et compte tenu du volume important d'information disponible sur le web, la récupération de toute l'information répondant à une requête de l'utilisateur est une tâche ardue. La disparité de l'information a pour effet d'augmenter le silence en recherche d'information sur le web.

1.2.4 La nature dynamique du web

En raison de la nature dynamique du web, les informations peuvent être ajoutées ou supprimées facilement. Il est estimé que 40% des pages web sont modifiées tous les mois. En outre, différents pages évoluent à des rythmes différents. Par exemple, les pages liées aux nouvelles (les médias), sports et les pages personnelles ont tendance à changer plus fréquemment que celles hébergées dans des domaines éducatifs ou gouvernementaux **[3].**

Cette nature dynamique du web rend la mise à jour et la maintenance des index des moteurs de recherche extrêmement difficile.

1.2.5 La fiabilité de l'information

L'information sur le web est produite par diverses sources ; cette diversité pose le problème, non moins crucial, de la qualité et de la fiabilité de l'information récupérée. En effet, il ne suffit pas de récupérer de l'information sur un sujet, encore faut-il savoir quelle valeur lui attribuer. L'information récupérée peut être une bonne information, une information non complète ou une information fausse ce qui est plus nuisible.

1.2.6 Les utilisateurs du web et leurs requêtes

Le plus souvent, les utilisateurs expriment leur besoin en information avec de petites requêtes, qui contiennent peu de mots, en moyenne 2.35 mots. Les courtes requêtes expriment d'une manière inexacte et ambiguë le besoin en information de l'utilisateur. En plus, la plupart des utilisateurs consultent seulement les premières pages retournées ; et s'engagent rarement dans le processus de la reformulation de la requête.

Ces caractéristiques du web rendent difficile pour les outils de recherche d'information actuels la tâche de sélection d'information désirée parmi le grand nombre de ressources qui répondent aux besoins des utilisateurs. Les limites des outils actuels de recherche d'information sur le web ont incité les chercheurs à développer de nouvelles approches pour aider à améliorer l'exactitude de la tâche de sélection de ressource.

1.2.7 La structure du Web

Le web peut être considéré comme un graphe orienté, dont les nœuds sont des pages web identifiées par des adresses URL et les arcs sont des hyperliens entre ces pages web. Trois approches ont été investies pour étudier la structure du web. La première est dite macroscopique, elle s'intéresse à la structure grande échelle du web, son but est d'avoir une vue de loin de graphe du web. La seconde approche est dite microscopique, sont objectif est

de rechercher dans le graphe du web de petites structures (ensemble de pages : communautés) qui se répètent souvent. Quant à la dernière approche, elle s'intéresse au calcul de certaines propriétés statistiques du web.

1.3 Les sources d'informations sur un document web

L'objectif majeur de la RI est le développement de stratégies qui permettent d'identifier tous les documents pertinents pour une requête de l'utilisateur. Dans la RI traditionnelle, seul le contenu textuel du document est considéré comme source d'information appropriée pour mesurer la pertinence d'un document vis-à-vis d'une requête. Dans le contexte d'un document web, d'autres sources d'information indépendantes du contenu textuel de document, peuvent être exploitées afin d'améliorer les performances de la RI. Ces sources d'information peuvent être : la structure du document, la structure des hyperliens.

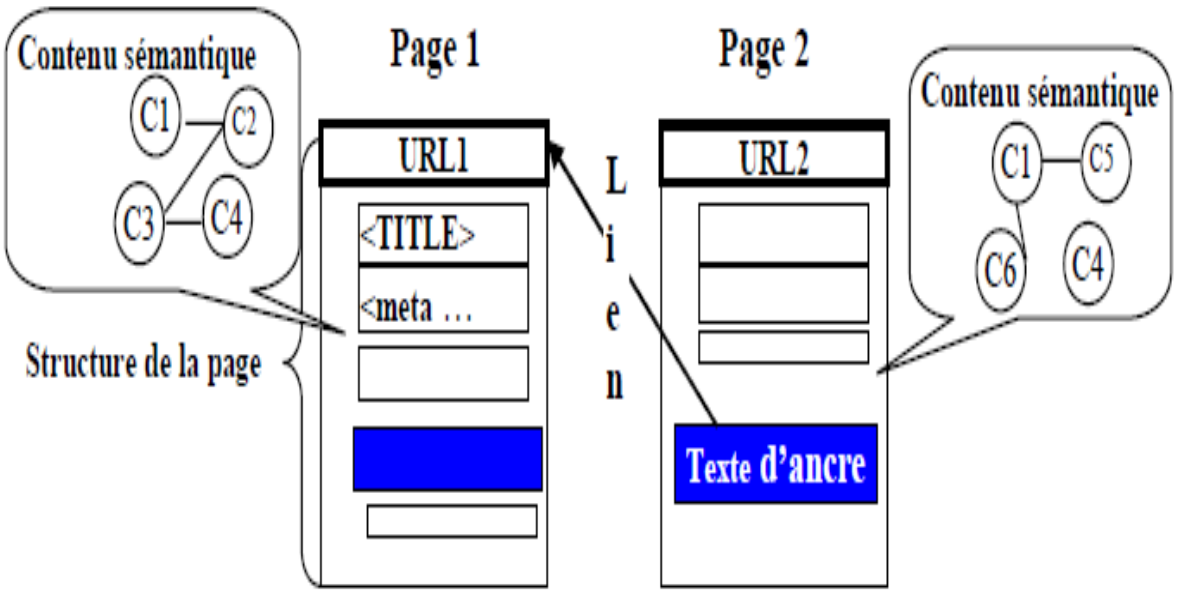


Figure 1.1 Les sources d'information sur un document web

1.3.1 Exploitation de la structure du document (page web)

Les SRI traditionnels négligent typiquement les informations sur la structure d'un document.

La raison principale est que généralement cette information est non disponible ou difficile à acquérir **[4].**

Il est reconnu que la pertinence de la recherche peut être améliorée en tenant compte de la structure d'un document. Plusieurs moteurs de recherche utilisent les balises de HTML pour améliorer la fonction d'appariement des documents. Les moteurs de recherche actuels (Google, Bing, Yahoo) donnent un bon score à un document qui contient les termes de la requête dans le titre de la page web.

Cutler *et al* **[4]** ont mené une étude sur l'apport de la structure des pages HTML pour améliorer la pertinence de la RI. La méthode d'indexation proposée consiste à associer à chaque terme un vecteur de fréquence noté *Tf v*, qui contient la fréquence d'apparition du terme dans une des classes de balises. Six classes de balises ont été utilisées : Anchor, <H1>-<H2>, <H3>-<H6>, STRONG, TITLE et le texte plein (toutes les autres balises).

Lors de la recherche un autre vecteur à six éléments noté *CIV* (Class Importance Vector) est utilisé, chaque élément de ce vecteur représente un facteur d'importance associé pour chaque classe de balises. L'importance (poids) d'un terme dans un document est alors calculée comme suit :

$$w = (Tf * CIV) * id\ f$$

Ainsi, la traditionnelle formule de pondération *Tf* est étendue comme suit : *Tf * CIV*, qui tient compte de la fréquence du terme dans une classe et de l'importance accordée à cette classe.

Les expérimentations menées ont montré que, l'usage de la structure des pages HTML améliore sensiblement la pertinence de la RI.

Olgivie *et al* **[5] [6] [7],** ont montré que la surpondération des termes apparaissant dans le texte des balises TITLE, ALT et FONT dans le cadre de la recherche de page d'entrée et de page nommée, n'apporte qu'un petit gain d'efficacité.

D'autre travaux ont utilisé les métadonnées (des données pour décrire les données sur lesquelles elles portent) en RI. Agosti et al **[8]** ont utilisé 15166 pages de la bibliothèque du Congrès[7], et ont observé que la recherche utilisant seulement le contenu a donné des résultats légèrement meilleurs que la recherche utilisant le contenu et les métadonnées. Cette étude est pe1u concluante car seulement une petite fraction des pages collectées contient des métadonnées.

Zhang et al **[9]** ont construit un ensemble de pages web (artificiellement, en ajoutant des métadonnées), et les ont soumis à un ensemble de moteurs, dans le but de mesurer

L'impact des métadonnées sur la visibilité de ces pages dans les moteurs de recherche.

Ils ont constaté qu'aucune des pages web construites qui contenant les termes des requêtes dans le champ de métadonnées ne figure dans les résultats de recherche. Ceci est dû vraisemblablement à l'utilisation des techniques anti-spam par ces moteurs de recherche.

En plus de la structure interne d'une page web, son adresse URL peut être une source d'information lors du calcul de la pertinence de la page, soit en considérant l'URL comme un texte, dans ce cas on peut appliquer les méthodes de recherche plein texte pour l'appariement entre les termes de la requête et le texte de l'URL, ou en considérant d'autres caractéristiques de l'URL comme : sa forme ou la présence de certains caractères, dans le but d'estimer la pertinence a priori d'un document (indépendamment de la requête).

Kraaij et al **[10]** ont utilisé la forme (type) de l'URL pour estimer la probabilité qu'une page soit une page d'entrée. Quant à Kamps et al **[11]**, ils proposent trois autres mesures, afin d'estimer la probabilité a priori de pertinence d'une page, en fonction du nombre de slash ('/') dans l'URL, du nombre de caractères de l'URL et enfin de la somme de nombre de caractères

'. ' Dans la partie domaine de l'URL et du nombre de '/' dans la partie chemin de l'URL. Les expérimentations menées ont montré que ces trois mesures sont de bons indicateurs afin d'estimer la probabilité a priori de pertinence d'une page.

1.3.2 Exploitation de la structure des hyperliens

L'analyse des liens a été étudiée avant l'apparition même du web, le domaine des réseaux sociaux s'y est intéressé pour diverses applications, comme la communication (détection d'espionnage, optimisation des transmissions). D'autres domaines se sont également penchés sur l'analyse des liens, c'est le cas de la bibliométrie, où l'analyse des références bibliographiques entre articles scientifiques est utilisée afin d'estimer le facteur d'impact des articles ou des journaux. D'autres mesures ont été également identifiées, la Co-citation et le couplage bibliographique.

Avec l'émergence du web, l'analyse de la structure des liens est au cœur de nombreux travaux en RI. Les deux principales utilisations des liens en recherche d'information concernent la collecte des pages web (Crawl) **[12]** et le classement des documents (Ranking). Les liens sont également utilisés pour d'autres fins comme : la classification et la catégorisation des pages web **[13],** la recherche des ressources dupliquées sur le web **[14],** la recherche de pages similaires **[15],** etc.

Nous nous intéressons particulièrement dans la suite de cette section à l'usage des liens dans la fonction de calcul de pertinence (classement des pages web).

Les méthodes de calcul de pertinence exploitant la structure des liens peuvent être réparties en trois classes distinctes, chacune d'entre elles est basée sur l'une de ces hypothèses suivantes **[16]**

1.3.2.1 L'hypothèse de recommandation :

Elle stipule que si une page est pointée par un lien alors cette page est recommandée par la page qui l'a référencé. Ainsi, une page avec beaucoup de liens entrants est une page fortement recommandée (populaire, autorité) et donc susceptible d'être mieux classée. Plusieurs algorithmes et méthodes de « Ranking » se sont basés sur cette hypothèse, pour le calcul de l'importance de la page. Nous décrivons ci-dessous quelques algorithmes représentatifs de cette classe.

- **Le nombre de liens**

Deux types de liens sont considérés pour une page web : les liens entrants et les liens sortants. Plusieurs études se sont penchées sur l'utilité du nombre de ces liens pour la recherche d'information. Dans **[17],** il est noté que le nombre de liens entrants peut fournir une indication sur l'importance, la popularité et la qualité de la page. Kamps et al **[11]** ont utilisé le nombre de liens entrants et sortants pour prédire l'importance de la page (probabilité a priori de pertinence) dans le cadre de la recherche de pages d'entrées et de pages nommées « Named-page ». Ils ont constaté que le nombre de liens entrants est un bon indicateur pour prédire la pertinence de la page.

- **PageRank**

Quelques moteurs de recherche, dont le plus connu est Google, ont pris le pari d'utiliser un autre mode de classement des résultats. Les pages Web sont ordonnées selon leur popularité, une page qui est la cible d'un très grand nombre de liens est probablement non seulement une page validée (page parcourue par un grand nombre de lecteurs, qui ont jugé bon de la citer en référence) mais aussi une page détenant un contenu utile à un grand nombre d'utilisateurs.

L'approche du PageRank qui a fait la spécificité du moteur de recherche Google, repose sur la notion de propagation de popularité. Le principe consiste à évaluer l'importance d'une page en fonction de chacune des pages pointant vers elle. La propagation met en avant les pages

qui jouent un rôle particulier dans le graphe des liens, avec l'hypothèse suivante : *"une page est importante quand elle est beaucoup citée ou citée par une page très importante*".

La mesure de PageRank (PR) est une distribution de probabilité sur les pages. Elle mesure en effet la probabilité PR, pour un utilisateur navigant au hasard, d'atteindre une page donnée. Elle repose sur un concept très simple : un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus cette page est considérée comme importante. Le PageRank se calcule de la façon suivante :

Ø Soient T1, T2, ..., Tn : n pages citant une page A. Notons PR(Tk) le PageRank de la

Page Tk, S(Tk) le nombre de liens sortants de la page Tk, et d'un facteur compris entre

0 et 1, fixé en général à 0.85. Ce facteur d représente la probabilité de suivre effectivement les liens pour atteindre la page A, tandis que (1-d) représente la probabilité d'atteindre la page A sans suivre de liens. Le PageRank de la page A se 1 calcule à partir du PageRank de toutes les pages Tk de la manière suivante :

$$PR(A) = (1 - d) + d\ (PR(T) / S(T)$$

Initialement, toutes les pages sont équiprobables, leur valeur de PR est alors égale à 1/*n*, n .Étant le nombre de documents de la collection.

- **HITS**

Klein berg a proposé un algorithme dit HITS (Hyper Link Induced Text Search) **[19]** pour identifier les documents autorités au moment de la recherche. L'algorithme est basé sur l'analyse de la matrice d'adjacence des pages retournées pour une requête donnée. Les pages web concernant le sujet de la requête peuvent être soit des Hub sou des Autorités. Les pages Autorités contiennent de l'information sur le sujet. Par contre, les pages Hubs pointent les pages Autorités. Ces deux types de pages ont une relation de renforcement mutuelle entre elles, ainsi : un bon Hub est une page qui pointe beaucoup de bonnes Autorités, et une bonne Autorité est une page qui est pointée par beaucoup de bons Hubs. La figure 1.2 ci-dessous illustre la relation entre deux types de pages **:**

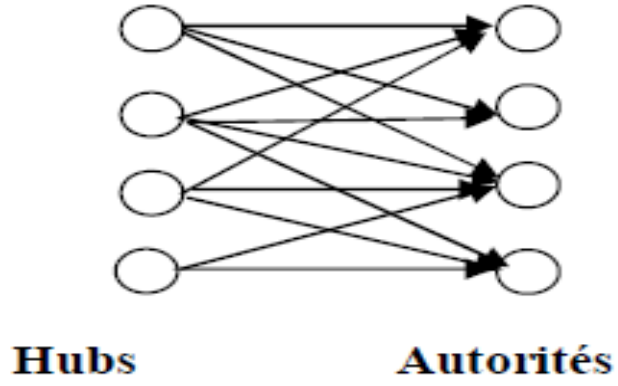


Figure 1.2 Les pages Hubs et Autorités

L'algorithme HITS calcule deux scores pour une page : le score d'Autorité et le score de Hub. L'évaluation de ces deux scores se fait au moment de l'interrogation, selon les deux formules ci-dessous, traduisant la relation de renforcement mutuelle entre les deux types de pages :

$$\mathbf{Ap} = \sum_{q \to p} \mathbf{H_q}$$

$$\mathbf{Hp} = \sum_{q \to p} \mathbf{A_q}$$

Où Ap et Hp représentent respectivement le score d'Autorité et de Hub de la page «p»,

$q \to p$ indique l'ensemble des pages qui pointent la page « p » $p \to q$ indique l'ensemble des pages pointées par la page « p ».

Le calcul des scores Ap et Hp se fait d'une manière itérative. Il se termine lorsque la convergence des scores s'opère, ces scores sont alors utilisés pour classer les documents. Plusieurs autres variantes ont été proposées élargissant les méthodes présentées précédemment, comme l'algorithme SALSA **[18].**

1.3.2.2 L'hypothèse de proximité sémantique

Elle stipule que si deux pages sont liées par un hyperlien, alors il est fort probable que les deux pages portent sur un même sujet, c'est-à-dire elles sont sémantiquement liées. Des travaux proposent alors de propager une fraction du score de pertinence d'une page vers ses pages voisines **[19] [20]** .De manière générale ces approches se basent sur deux étapes : la première consiste à classer les pages web selon la similarité de leur contenu avec la requête.

Dans la seconde étape, les premières pages seulement sélectionnées précédemment transmettent une fraction de leur score vers leurs voisines. L'orientation des liens peut être respectée ou ignorée. Ainsi, un nouveau score est calculé pour chaque page selon la formule suivante :

$$\mathbf{RSV\ (d_i, q) = Sim\ (d_i, q) + \lambda \sum_{j=1}^{m} Sim\ (d_i, q)}$$

Où **RSV (d_i, q)** représente le nouveau score du document **d_i** dépendant de son score initial noté Sim (d_i, q) et du score de ses **m** voisins et **λ** est un facteur de pondération. Savoy et al **[20]** ont noté que les différentes évaluations effectuées sur un corpus d'articles et sur une collection de pages extraites du web n'ont pas amélioré d'une manière significative les résultats, en appliquant différentes valeurs pour le facteur O. Cependant, ils ont constaté une amélioration en ne considérant que les pages possédant un meilleur score initial.

1.3.2.3 L'hypothèse de la description du texte d'ancre

Un texte d'ancre donne une petite description de la page référencée. Ceci le rend utile pour l'indexation de la page pointée plutôt que la page source. Une manière simple d'exploiter cette source est de prendre en compte tous les textes d'ancre pointant un document, au même titre que le contenu de ce document lors du calcul de son score de pertinence. Dès les années 90, des moteurs de recherche tels que le moteur Altavista ont recouru à l'utilisation du texte d'ancre ; et il a été montré qu'il est un élément utile pour la RI sur le web **[21].**

Des études académiques ont rejoint cette constatation, et cela dans le cadre de la recherche de pages d'entrée et de la recherche du site **[16] [10]**.

Plusieurs facteurs (la plupart statistiques) laissent à penser que le texte d'ancre est utile pour l'amélioration de la qualité de la RI **[22]:** les requêtes des utilisateurs sont dans la plupart des cas de petites tailles, cette caractéristique est généralement partagée par les textes d'ancre; la requête comme le texte d'ancre tendent à ne pas être des phrases complètes, et leurs vocabulaires ainsi que leurs formes grammaticales sont semblables; les termes de la requête et les termes du texte d'ancre appartiennent généralement au même espace concept (décrivent le même concept). En plus de ces facteurs liés à la requête, d'autres facteurs liés au document plaident pour l'usage des textes d'ancre : il est fréquemment observé que le texte d'ancre contient des termes qui n'apparaissent pas dans le contenu de la page pointée, donc le texte d'ancre est une source d'information supplémentaire pour le contenu de la page ; des points de vue différents sur un même contenu sont donnés par les auteurs des pages qui pointent une

page donnée. Ce qui est reproché à l'usage des textes d'ancre ; est que certaines pratiques de la part des auteurs des pages peuvent avoir un impact négatif sur le résultat de la recherche. Par exemple, pour la requête « more evil than Satan himself », le moteur Google qui utilise le texte d'ancre proposait, en octobre 1999, le site web de Microsoft comme réponse la plus pertinente. Cette réponse est la conséquence de l'existence des termes de la requête dans les textes d'ancre (afin de dénoncer les pratiques commerciales de Microsoft) qui pointent le site web Microsoft. Les textes d'ancre peuvent aussi contenir des termes peu informatifs, exemple (cliquez ici, Bas, Haut, etc.). Pour contourner ce problème et prendre en compte le contexte du texte d'ancre plusieurs travaux ont proposé d'inclure le texte encadrant le texte d'ancre. Ainsi, Chakrabarti et al **[23]** ont examiné la distribution du terme "Yahoo" autour d'ancre de http://www.yahoo.comdans 5000 pages. Ils ont trouvé que la plupart des occurrences de ce terme font partie d'un cadre de 50 termes, et ont montré que l'usage du texte de proximité améliore le rappel en dépit de la précision.

Glover et al **[24]** quant à eux ont montré que l'usage d'un cadre de 25 termes autour du texte d'ancre améliore les performances de classification des pages web.

D'autres sources d'information sur un document web ont étés utilisées, comme le facteur temps **[25] [26]** et le rapport information/bruit **[17].**

1.4 La combinaison des sources d'information

Il est reconnu que la combinaison de différentes sources d'information sur un document peut améliorer l'efficacité d'un SRI **[27]** .Cette combinaison est réalisée suivant deux approches:

- **Combinaison de résultats :** elle consiste à traiter chaque source d'information (ou plusieurs) comme un SRI à part, chacun de ces systèmes peut utiliser différentes stratégies de recherche; puis à combiner les résultats de recherche obtenus par ces derniers dans une seule liste ordonnée. Cette approche se réfère à la fusion de données dans les environnements de RI traditionnels et à la méta-recherche dans le contexte du web.
- **Combinaison de facteurs :** elle consiste à développer des modèles (ou étendre les modèles existants) qui supporteront la combinaison de plusieurs sources d'information sur un document sous un même cadre. Souvent, ce sont des modèles basés sur l'apprentissage automatique.

1.4.1 Combinaison de résultats

Plusieurs stratégies de combinaison de résultats de recherche de différents SRI ont été proposées. Certaines se basent sur les scores des documents retournés par chaque système pour effectuer la fusion et d'autres se basent sur le rang des documents dans les listes retournées par chaque SRI. Le tableau 1.1 recense les différentes méthodes de combinaison de résultats.

Méthodes bassées sur le rang	Borda-fuse et Borda-fuse pondérée**[28]** Ré ordonnancement **[29]**
Méthodes bassées sur le score	CombMNZ, CombSUM, CombMAX, CombMIN et CombMED **[30]** ; Combinaison linéaire **[31]**

Tableau 1.1 Stratégies de combinaison de résultats

1.4.2 Combinaison de facteurs

Cette approche de combinaison consiste à développer des modèles de RI qui supporteront explicitement la combinaison de plusieurs sources d'informations sur un document, sous un cadre unique.

Nous citons ci-dessous quelques travaux ayant été réalisés dans ce sens. En 1988, Fox et al **[32]** est mené un certain nombre d'expérimentations sur ce type de combinaison dans le cadre du modèle vectoriel. Ils ont proposé de décrire chaque représentation d'un document par un sous-vecteur. Par exemple : un sous-vecteur pour les termes, un autre pour les auteurs et un autre pour les citations. La fonction de similarité document-requête est alors une combinaison linéaire des similarités des différents sous vecteurs avec la requête.

Tsikrika et al **[33]** est utilisé le modèle de réseau bayésien pour la combinaison des différentes représentations d'un document web. Le modèle proposé est composé de quatre couches : couche documents, couche représentations des documents, couche requêtes et couche besoins des utilisateurs. Les deux premières couches sont construites au moment de l'indexation et les deux dernières au moment de l'interrogation. Deux sources d'information sont explorées, le contenu du document, à partir de laquelle une représentation du document est construite (représentation du contenu) et la structure d'hyperliens, pour laquelle deux types de

représentation sont construites, le texte (étendu) d'ancre des liens entrants et le texte (étendu) des liens sortants de la page. Sur la base du typage sémantique des liens (trois types de liens : composition, séquence et référence), huit représentations au total sont obtenues. En plus des textes d'ancre, l'utilisation de l'algorithme HITS a été aussi investie.

Les expérimentations menées en utilisant la collection TREC WT2g ont montré que :

- Le contenu de la page et le texte d'ancre des liens entrants donne de meilleurs résultats que ceux obtenus avec le texte des liens sortants.
- Les résultats de la combinaison des différentes sources d'information ont montré que le contenu de la page est la source d'information la plus importante lors de la combinaison. De plus, une représentation qui donne des résultats individuels faibles peut améliorer les performances du système en la combinant avec d'autres représentations.
- L'application de l'algorithme HITS n'améliore pas les résultats obtenus avec le contenu de la page.

Dans le cadre du modèle de langue, plusieurs travaux ont été proposés pour combiner différentes sources d'information sur un document web **[25] [10] [26] [7] [17] .** Ce point est discuté en détails dans le chapitre suivant (section 3.5).

D'autres modèles basés sur des algorithmes d'apprentissage automatique (en anglais Learning to Rank) ont été proposés **[34] [35].**

L'idée de base de ces algorithmes est de faire apprendre une fonction d'ordonnancement en assignant un poids pour chaque source d'information sur un document, puis utiliser la fonction obtenue pour estimer le score de pertinence de chaque document, et en fin ordonner les documents selon leur score de pertinence obtenu. On distingue trois grandes approches d'algorithmes d'ordonnancement : par point (pointwise), par paire (pairwise) et par liste (listwise). Ces approches diffèrent sur leur façon de considérer le problème d'apprentissage **[36]**.

L'approche par point (pointwise) considère les documents séparément en entrée du système d'apprentissage. A chaque document est associé un score (ou un degré) de pertinence pour une requête donnée. Le problème d'apprentissage est alors assimilé à un problème de régression **[37]** ou de classification **[38]** respectivement.

L'approche par paire (pairwise) considère en entrée du système d'apprentissage des paires de documents (di, dj) auxquels sont associées des jugements de préférence $\mathbf{r_{i,j}}$ à valeur dans {1,1}. Si $\mathbf{r_{i,j} = 1}$ alors le document di est préféré au document dj, il doit être mieux classé dans la liste de résultat. La préférence est notée $d_i > dj$. Au contraire, si $\mathbf{r_{i,j} = -1}$ alors le dj est préféré au document di et on note $dj > d_i$. Le problème d'apprentissage est ici un problème de classification, dans le cas particulier de paires d'instances. Plusieurs techniques ont été proposées pour le classement des documents **[39[40].**

Enfin, l'approche par liste (listwise) considère en entrée du système d'apprentissage une liste ordonnée de documents. La fonction d'ordonnancement est apprise par minimisation de la distance entre la liste apprise et la liste de référence **[41]** ou par optimisation d'une mesure de recherche d'information **[42]**.

1.5 Conclusion

Nous avons abordé dans ce chapitre la RI sur le web. Particulièrement, les points suivants ont été étudiés. En premier lieu, nous avons énuméré les éléments distinctifs entre la RI classique et la RI sur le web. Ensuite, nous avons identifié et étudié les différentes sources d'information d'un document web. Enfin, nous avons présenté les différentes approches (méthodes) proposées pour combiner ces différentes sources d'information. Parmi ces méthodes, nous nous intéressons à celles basées sur l'utilisation (exploitation) d'un cadre unique pour la combinaison des sources d'information sur un document web. Plus explicitement, nous utilisons dans notre travail, le modèle de langage comme cadre de combinaison de ces sources d'information. Dans le chapitre suivant nous détaillons ce modèle, ainsi que les travaux réalisés dans ce cadre pour intitule Ingénierie Ontologique.

Chapitre 2

Ingénieries des ontologies

2. Introduction

Durant ces dernières années, les ontologies sont largement utilisées et ont prouvé leurs utilités dans de nombreux domaines tels que : l'ingénierie de connaissances, l'intelligence artificielle, l'intégration des sources de données, la recherche d'information, le commerce ´électronique et sont au cœur du Web sémantique **[T. Berners-Lee, Hendler & O. Lassila, 2001]** Cet engouement est motivé par le fait que les ontologies sont un moyen efficace pour la gestion et le partage des connaissances d'un domaine particulier entre personnes et/ou systèmes. Ce chapitre a pour titre ingénierie ontologique qui est une discipline de l'informatique référant à l'ensemble des activités qui concernent le processus de développement d'ontologie, les méthodes et les cadres méthodologiques de sa construction, donc Nous allons présenter dans ce chapitre, en premier lieu la notion d'ontologie, en relevant quelques définitions qui lui ont été attribuées et quelques intérêt plus les différents éléments qui la constituent. Par la suite, nous présenterons ses différentes classifications en se concentrant sur la richesse de la structure interne de l'ontologie. Ainsi, on va désigner les différentes méthodologies de conception. En outre, nous faisons un survol des principaux formalismes de représentation de connaissances à savoir les frames, les logiques de description et les réseaux sémantiques qui sont à l'origine des langages permettant d'exprimer des ontologies.

2.1 Définitions d'une ontologie

Le terme ontologie vient du domaine de la philosophie au XIXème siècle, qui s'intéresse à l'étude de l'être ou de l'existence. En philosophie, l'ontologie est une branche de la métaphysique qui s'intéresse à la notion d'existence, son premier sens a trouvé son origine depuis ARISTOTE, ou` l'ontologie est l'´étude des propriétés générales de ce qui existe. Par analogie, le terme est repris en informatique et sciences de l'information dans les années 80, ou l'ontologie est un terme technique désignant un artefact **[Tom Gruber, 2009]** qui est conçu pour un but, qui est de permettre la modélisation des connaissances sur un domaine quelconque, réel ou imaginaire. Les ontologies sont employées dans l'intelligence artificielle, le Web s´sémantique, le génie logiciel, l'informatique biomédicale ou encore l'architecture de l'information comme une forme de représentation de la connaissance au sujet

d’un monde ou d’une certaine partie de ce monde. Les ontologies informatiques sont des outils qui permettent précisément de représenter un corpus de connaissances sous une forme utilisable par un ordinateur.

En informatique, plusieurs d´définitions ont été données pour l’ontologie :

En 1991, Neches et ses collègues ont été les premiers à en proposer une définition :

Une ontologie définit les termes et les relations de base du vocabulaire d’un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire

En 1993, Gruber a proposé une définition a` cette notion qui est la plus célèbre couramment citée dans la littérature :
Une ontologie est une spécification explicite d’une conceptualisation.

En 1997, Borst a modifié légèrement la définition de Gruber en citant qu’une ontologie est définie comme ´étant :
Une ontologie est une spécification formelle d’une conceptualisation partagée.

En 1998, Studer et ses collègues ont rassemblé ces deux définitions (celles de Gruber et Borst) dans une seule qui est :

Une ontologie est une spécification formelle et explicite d’une conceptualisation partagée.

- Spécification explicite : signifie que les concepts, les propriétés, les relations, les fonctions, les restrictions et les axiomes de l’ontologie sont d´définis de façon déclarative.
  - Formelle : réfère au fait qu’une ontologie doit être traduite dans un langage interprétable par une machine.

- Conceptualisation **:** réfère à un modèle abstrait d'un phénomène du monde en identifiant les concepts appropriés à ce domaine.
- Partagé : réfère au fait qu'une ontologie capture la connaissance consensuelle c'est-à-dire non réservée à quelques individus, mais partagée par un groupe ou une communauté. **[F. Amourache, 2008]**

C'est une base de formalisation des connaissances. Elle se situe à un certain niveau d'abstraction et dans un contexte particulier.

C'est aussi une représentation d'une conceptualisation partagée et consensuelle, dans un domaine particulier et vers un objectif commun. Elle classifie en catégories les relations entre les concepts.

2.2 Intérêts de l'ontologie

Quelques intérêts de l'utilisation des ontologies :

- Vocabulaire unifié : un vocabulaire commun constitue le rôle d'une ontologie, qui permet l'unification des vocabulaires d'un domaine particulier.
- Structuration des concepts : la structuration des données d'un domaine par une hiérarchie des concepts.
- Interopérabilité : l'ontologie constitue une solution intéressante pour le problème d'interopérabilité entre les systèmes d'informations hétérogène. Elle permet de d´définir clairement les différents concepts utilises en vue d'assurer une communication sans ambigüité entre les systèmes hétérogène.
- La communication : elle peut avoir lieu entre les hommes et/ou les systèmes. Les ontologies permettent alors le partage de la compréhension et la communication dans des contextes particuliers et selon les besoins. Ainsi, on peut utiliser l'ontologie pour créer un réseau de relations qui définit les connexions entre les composants du système. Cette caractéristique de communication est offerte grâce à la non-ambigüité des termes utilises et définis par l'ontologie dans le système.
- Explication de l'implicite **:** l'ontologie explicite les connaissances implicite pour résoudre le problème d'ambigüité entre concepts.

- Proposition d'un méta- modèle : par définition, un modèle constitue une abstraction d'une réalité. Comme la notion d'ontologie se base sur la définition des concepts et des relations elle s'intègre parfaitement dans une solution pour création des modèles.

2.3 Les composants d'une ontologie

2.3.1 Concepts

Un concept peut représenter un objet matériel, une notion, une idée; les concepts sont aussi appelés termes ou classe de l'ontologie, constituent les objets de base manipulés par les ontologies. Ils correspondent aux abstractions pertinentes du domaine du problème, retenues en fonction des objectifs qu'on se donne et de l'application envisagée pour l'ontologie. **[M. Uschold & M. King ,1995]**.

Selon **[Gomez-Perez, 1999]**, les concepts peuvent être classifiés selon plusieurs dimensions :

- Niveau d'abstraction : concret ou abstrait,
- Atomicité : élémentaire ou composé,
- Niveau de réalité : réel ou fictif,
- Un concept peut aussi être divisé en trois parties :
- Un Terme : ce dernier permet de désigner un concept ou bien 'label de concept'. **Exemple :** Médecin
- Une Intension : ou notion est un ensemble de propriétés qualitatives ou fonctionnelles communes aux individus auxquels le concept s'applique, et permettant de d´définir le concept.

  Exemple : Etablissement de soin conçu et aménagé pour les soins d'un grand´ nombre de personnes malade.

- Une Extension: ou ensemble d'objets qui regroupe les objets manipulés à travers le concept, autrement dit instance du concept.

  Exemple : La liste des établissements de soin

Par ailleurs, un terme peut ne pas avoir une extension comme par exemple le concept Vérité qui a le sens de Ce Qui Est Vrai, il s'agit alors d'un concept générique qui correspond à une notion abstraite. **[G.Falquet ,01]**

Souligne qu'on peut trouver des concepts partageant la même extension mais pas leur intention et portent le même terme, ceci correspond à des points de vue différents sur un même concept.

Un Concept peut être caractérisé selon les propriétés qui lui sont associé, comme suit :

- **La généricité :** un concept est dit générique s'il n'admet pas d'extension, **Exemple :** La vérité est un concept générique.

- **L'identité :** un concept porte une propriété d'identité si cette dernière peut différencier deux instances de ce concept, par **exemple :** le concept étudiant porte une propriété d'identité liée au numéro de l'étudiant, deux étudiants ne peuvent pas avoir le même numéro.
- **La rigidité :** un concept est dit rigide si toute extension de ce concept reste une extension de toutes les connaissances possibles, **Exemple :** Humain concept rigide. Etudiant concept non rigide.
- **L'anti-rigidité :** un concept est dit anti-rigide s'il peut être une instance pour d'autres concepts, par **exemple :** Chercheur est un concept anti rigide car Chercheur est avant tout Humain.
- **L'Unité** : un concept est dit concept unité, si pour chacune de ses instances, les différentes parties de l'instance sont liés par une relation qui ne lie pas d'autres instances de concepts, **par exemple :** les deux parties d'un couteau sont reliés par une relation Lié, poignée est liée à une lame.

Deux concepts portent les propriétés suivantes :

- **L'´équivalence** : Deux concepts sont équivalents s'ils ont la même extension.
- **La disjonction :** Deux concepts sont disjoints si leurs extensions sont disjointes, **exemple :** Lumière, Obscurité.

- **La dépendance** : Un concept C1 est dépendant d'un concept C2 si pour toute instance de C1 il existe une instance C2 qui ne soit ni partie ni constituant de l'instance C2, **par exemple**: Père est un concept d´dépendant de Fils et vice-versa.

2.3.2 Relations

Les relations traduisent les interactions existant entre les concepts présents dans le domaine d'analyse. Elles sont formellement définies comme tout sous ensemble d'un produit cartésien de n ensembles, c'est-à-dire **R : C1 x C2x ... x Cn**

Les concepts peuvent être reliés entre eux par des relations au sein d'une ontologie. **Par exemple :** la relation Ecrit lie une instance de concept Personne et une instance du concept Article, dans cet ordre. Ces relations incluent les associations suivantes :

- Sous-classe-de/Is-a : Spécialisation, généralisation,
- Partie-de : agrégation ou composition,
- Associé-a, instance-de Ces relations nous permettent de capturer la structuration ainsi que l'interaction entre les concepts, ce qui permet de représenter une grande partie de la sémantique via l'ontologie

2.3.3 Fonctions

Les fonctions sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 éléments précédents. Formellement, les fonctions sont définies ainsi :

$$\mathbf{F : C1xC2...XCn - 1 \rightarrow Cn.}$$

2.3.4 Axiomes

Permettent de modéliser des assertions toujours vraies, à propos des abstractions du domaine traduites par l'ontologie. Ils permettent de combiner des concepts, des relations et des fonctions pour définir des règles d'inférences et qui peuvent intervenir, par exemple, dans la déduction, la définition des concepts et des relations, ou alors pour restreindre les valeurs des propriétés ou les arguments d'une relation.

2.3.5 Les instances

Les instances ou individus constituent la définition extensionnelle de l'ontologie. Ils représentent des éléments singuliers véhiculant les connaissances à propos du domaine du problème.

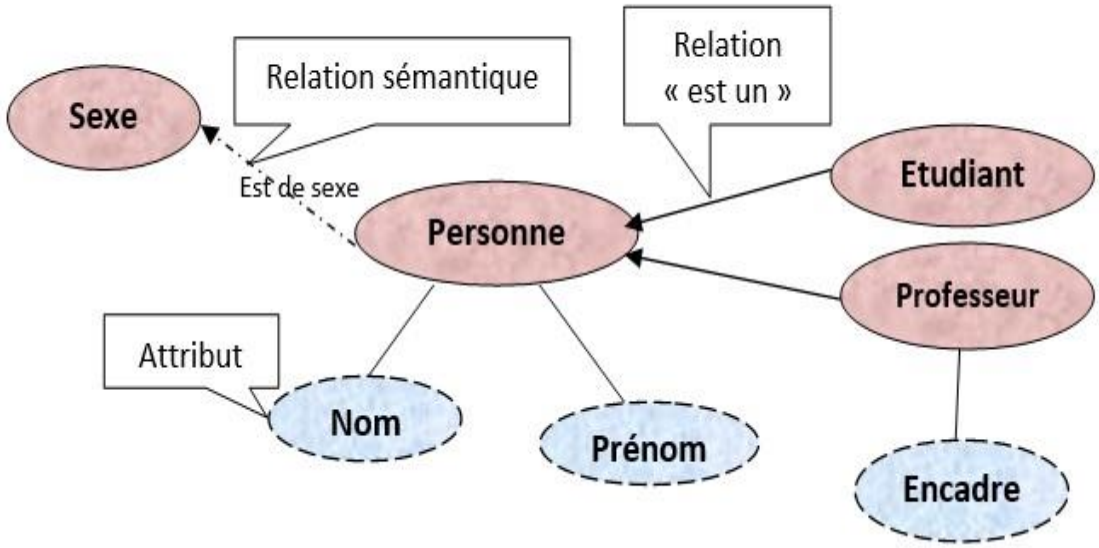


Figure 2.1 schéma qui représente les déférentes structures d'ontologie

2.4 La Typologie des ontologies

Les ontologies peuvent être classifiées selon plusieurs dimensions à savoir : objet de conceptualisation, niveau de formalisme de représentation, niveau de détail, et niveau de complétude. **[Psyché, Mendes & Bourdeau, 2004]** Comme le proposent **[N. Guarino, 1998]** la classification peut également se faire en fonction des objets que modélisent les ontologies pour répondre à un objectif précis :

2.4.1 Typologie selon l'objet de conceptualisation

- **Les ontologies de représentations des connaissances**

Ce type d'ontologies regroupe les concepts impliqués dans la formalisation des connaissances. On peut citer l'exemple de l'ontologie de frames **[Gomez- Perez, 1999]**,

qui définit les primitives de représentation des langages à base de frames : Classes, Instances, facettes, Propriétés, relations.

- **Les ontologies de niveau supérieures (Haut niveau)**

Ce type d'ontologies vise à étudier les catégories des choses qui existent dans le monde, comme les concepts de niveau supérieur d'abstraction tels que les entités, les événements, les processus, les actions, le temps, les relations et les propriétés. L'ontologie de niveau

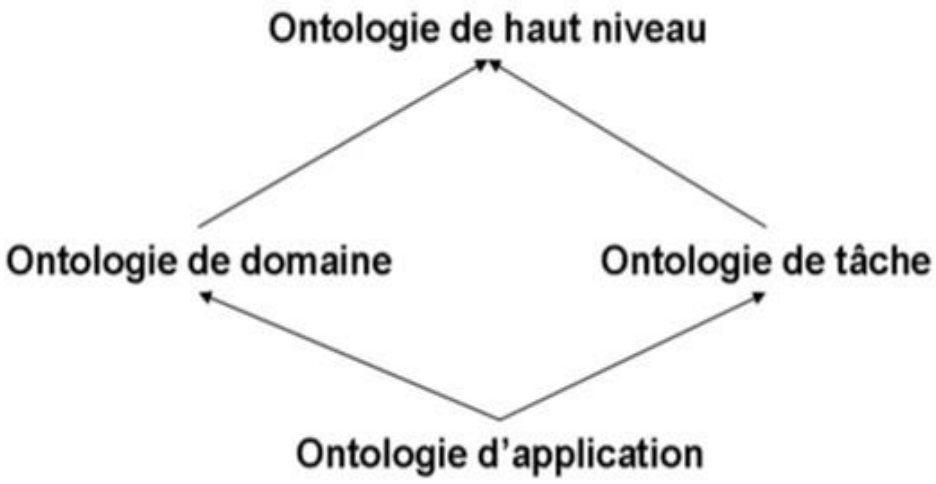


Figure 2.2 Classification des Ontologies selon **[N.Guarino, 1998]**

Supérieur est Fondé sur la théorie de l'identité et la théorie de la dépendance. Des recherches ont été poursuivies sur la théorie de l'ontologie, ces dernières ont affirmé que les fondements philosophiques étaient des principes à suivre afin de concevoir l'ontologie de niveau supérieur. **[Guarino, 1997]**

- **Les ontologies génériques**

Ce type d'ontologies aussi appelée méta-ontologies, véhiculent des connaissances génériques moins abstraites que celles véhiculées par l'ontologie de niveau supérieure, ces dernières peuvent être réutilisées à travers différents domaines. Elle peut adresser des connaissances réelles ou encore des connaissances visant à résoudre des problèmes génériques à travers différents domaines. **[Gomez &Perez, 1999]**

On pend deux exemples d'ontologies citées par **[Borst, 1997]** :

- l'ontologie m´érotologique qui est une branche de l'ontologie formelle et qui est une application de la logique des prédicats qui traite des relations entre la partie et le tout comme par **exemple :**

  Le Toit d'une maison, Les organes du corps, Les cellules de l'organisme,

- l'ontologie topologique tout comme la théorie de m´érotologie, l'ontologie topologique contient des relations, associé-a`.

- **Les Ontologies de domaine**

Ce type d'ontologies est construit sur un domaine de connaissances bien spécifiques; elles fournissent le vocabulaire des concepts de ce domaine et les relations entre ces derniers, les activités de ce domaine ainsi que les théories et les principes de base de ce domaine. Ces ontologies de domaine constituent donc des méta-descriptions d'une représentation de connaissances du domaine. La plupart des ontologies existantes sont des ontologies du domaine. Elle caractérise la connaissance du domaine ou` la tâche est réalisée. Un domaine serait par exemple : la médecine, la recherche.

- **Les Ontologies de tâches**

Ce type d'ontologies est utilisé pour gérer ou conceptualiser des tâches spécifiques dans les systèmes, telles que les tâches de diagnostic, de planification, de conception, de configuration et de tutorat, soit tout ce qui concerne la résolution de problèmes. Elle fournit un ensemble de vocabulaires et de concepts qui décrit une structure de résolution des problèmes inhérente aux tâches et indépendante du domaine.

- **Les Ontologies d'application**

Ce sont les ontologies les plus spécifiques. Elles permettent de décrire des concepts dépendants à la fois d'un domaine et d'une tâche. Dans cette classification, la notion d'ontologie d'application définit le contexte d'une application qui décrit la s´sémantique des informations et des services manipulés par une ou un ensemble d'applications sur un même domaine.

2.4.2 Selon le niveau de formalisme de représentation

Ceci est lié au niveau du formalisme de représentation du langage utilisé pour rendre l'ontologie opérationnelle. Nous pouvons citer quatre catégories qui sont :

- **Ontologies Informelles**

Ontologies opérationnelles dans un langage naturel sans aucune restriction (sémantique ouverte), cela aide à rendre l'ontologie plus compréhensible pour l'utilisateur.

- **Ontologies Semi-informelles**

Utilisation d'un langage naturel structuré, cela permet d'augmenter la clarté de l'ontologie tout en réduisant l'ambigüité.

- **Ontologies Semi-formelles**

Utilisation d'un langage artificiel défini formellement en utilisant des classes, des relations, des fonctions, objets et des axiomes pour décrire une ontologie.

- **Ontologies Formelles**

Utilisation d'un langage artificiel contenant une sémantique formelle, ainsi que des théorèmes et des preuves des propriétés telles la robustesse et l'exhaustivité. **[Gomez-Perez, 1999]**

2.4.3 La typologie selon le niveau de détail

Par rapport au niveau de détail utilisé lors de la conceptualisation de l'ontologie en fonction de l'objectif opérationnel envisagé pour l'ontologie, deux catégories au moins peuvent être identifiées :

- **Granularité fine**

Correspondant à des ontologies très détaillées, possédant ainsi un vocabulaire plus riche capable d'assurer une description détaillée des concepts pertinents d'un domaine ou d'une tâche. Ce niveau de granularité est très utile afin d'établir un accord entre les agents qui utiliseront cette ontologie.

- **Granularité large**

Correspondant à un vocabulaire moins détaillé, par exemple l'ontologie de haut niveau possède une granularité large; les concepts qu'elle traduit sont normalement détaillés dans d'autres ontologies de domaine ou d'application.

2.5 Méthodologie de construction d'une ontologie

Il existe plusieurs méthodes d'ingénierie ontologique mais jusqu'à présent aucune d'elles n'est proposée comme une méthodologie générale de construction d'ontologie ou bien normalisée, de ce faite certains auteurs ont tenues à proposer des méthodologies inspirées de leur expérience de construction d'ontologie.

On entend par méthodologie un ensemble de procédures de travail, d'étapes, un cycle de développement qui pourra être adopté lors de la construction d'ontologies C'est dans les années 1995 que les premières ontologies commencent à naître et parmi ces ontologies on cite :

2.5.1 La méthode ENTREPRISE

Ils ont proposés une première méthode de construction d'ontologie inspirée de leur expérience acquise lors du développement des ontologies dans le domaine de la gestion **[M. Uschold & M. King, 1995]**, Cette méthode repose sur les quatre étapes suivantes :

- identification du but et la portée de l'ontologie dont les raisons pour lesquelles l'ontologie en cours de construction est clarifiée ainsi que les utilisateurs potentiels de l'ontologie.
- Construire l'ontologie cette étape est divisée en trois activités :
- **Capture de l'ontologie** : Identifier les concepts et les relations fondamentaux, produire des définitions précises et non ambigües à ces éléments en langage naturel, identifier les termes dénotant ces éléments et enfin essayer d'arriver à un argument.

Désormais, cette analyse est insuffisante afin de spécifier la sémantique de ce domaine, certaines notions doivent d'abord être lues par un expert ou un

spécialiste du domaine. La sémantique doit alors être validée par les experts du domaine.

- **Approche descendante**, à partir d'un nombre réduits de concepts qu'on veut spécialiser. **Approche ascendante**, consiste à` considérer les termes spécifiques et de trouver des termes génériques associés.
- **Approche intermédiaire**, dans laquelle les concepts se structurent autour des concepts importants du domaine ni trop généraux, ni trop spécifiques.

- **Codage de l'ontologie :** La représentation explicite de la conceptualisation dans un langage formel.

- **Réutiliser et intégrer :** Eventuellement des ontologies existantes, Cette´ activité peut être effectuée en parallèle avec l'activité de capture et/ou de codage.

➢ ´évaluer l'ontologie. Cette ´étape consiste à confronter l'ontologie aux objectifs, logiciels et utilisateurs pour lesquelles elle a été élaborée, évaluer l'ontologie en vérifiant la validité de la taxonomie et vérifier si toutes les instances d'une classe sont aussi des instances de la classe mère et qu'il n y a pas de classe isolée.

➢ documenter l'ontologie. Cette étape consiste à établir des instructions de documentation d'ontologie qui diffèrent selon le type et le but.

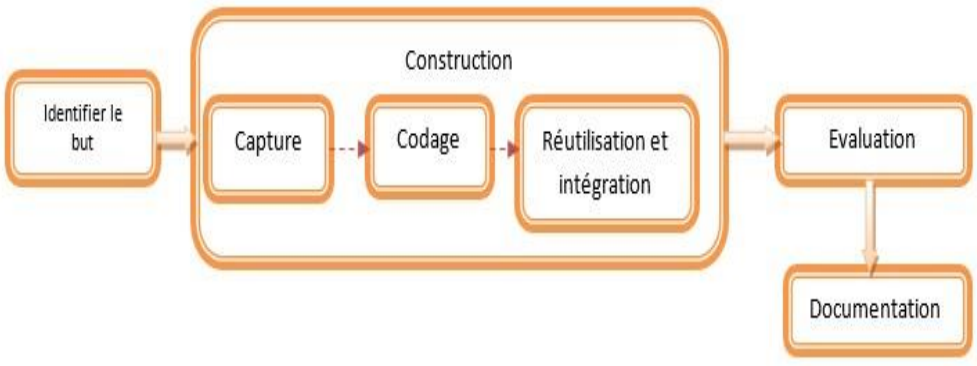


Figure 2.3 méthode d'Uschold et King.

2.5.2 La Méthode TOVE

Cette méthode est basée sur l'expérience du développement de l'ontologie du projet TOVE (Torento Virtual Enterprise). Elle aboutit à la construction d'un modèle logique de connaissance. **[M.gruninger & M. fox, 1995]** l'ontologie est développée selon les étapes suivantes :

- Identification des scénarios Cette étape consiste à identifier les applications possibles dans laquelle l'ontologie sera employée.
- Formulation de questions informelles Cette étape consiste à formuler un ensemble de questions basées sur des scénarios exprimées en langage naturel, afin de déterminer la portée de l'ontologie. Ces questions et réponses sont utiles afin d'extraire les concepts principaux, leurs propriétés et les relations qui existe entre ces concepts.
- Spécification Formelle Cette étape consiste à représenter des axiomes et des définitions pour les termes de la terminologie, en utilisant le formalisme de la logique du premier ordre. Les concepts seront représentés sous forme de constantes ou bien de variables. D'autre part, les propriétés et les relations seront représentées par des prédicats.
- Evaluation de la complétude de l'ontologie Cette étape consiste à effectueŕ deux taches telles que la vérification et la validation, en assurant la conformité et la fidélité sémantique de l'ontologie au domaine de connaissances.

2.5.3 La Méthode METHONTOLOGY

Cette méthode est développée au Laboratoire d'Intelligence Artificielle de l'Université polytechnique de Madrid. Elle vise la construction d'ontologies au "niveau connaissance" **[Fernandez-Lopez, 1999]** et repose sur :

- **Spécification**

  Cette étape a pour but de fournir une description claire du problème étudié ainsi que la façon de le résoudre. Elle permet de préciser l'objectif, la portée et le degré de granularité de l'ontologie qui sera construite.

- **Conceptualisation**

  Cette étape consiste à organiser et structurer la connaissance acquise durant la phase spécification en utilisant des langages d'implémentation dans lesquelles l'ontologie va être formalisée. Les représentations intermédiaires utilisées sont : les taxonomies de concepts, les diagrammes des relations binaires, le glossaire des termes, le dictionnaire des concepts, le tableau des relations binaires, spécifier des contraintes sur les attributs dans une table d'attributs, spécifier des axiomes sur les concepts dans une table d'axiomes logiques, décrire les instances des concepts dans une table d'instances.

- **Formalisation**

  Cette étape consiste à choisir ou à construire un langage formel capable d'intégrer toutes les propriétés de l'ontologie l'intérêt de l'utilisation d'un langage formel et de permettre de réduire les ambigüités du langage naturel en offrant une plus grande expressivité et d'autre part rendre l'ontologie compréhensible par les machines.

- **Implémentation**

  Cette étape consiste à la codification de l'ontologie formelle dans un langage opérationnel du Web Sémantique.

- **Maintenance**

  Cela peut s'agir d'une maintenance évolutive de l'ontologie (nouveaux besoins de l'utilisateur), ce qui permet la validation de celle-ci. Cette activité est généralement faite par le constructeur et des experts du domaine. La validation se base sur l'exploitation des services d'inférences associés a` la logique de Description, et qui sont offerts par des raisonneurs.

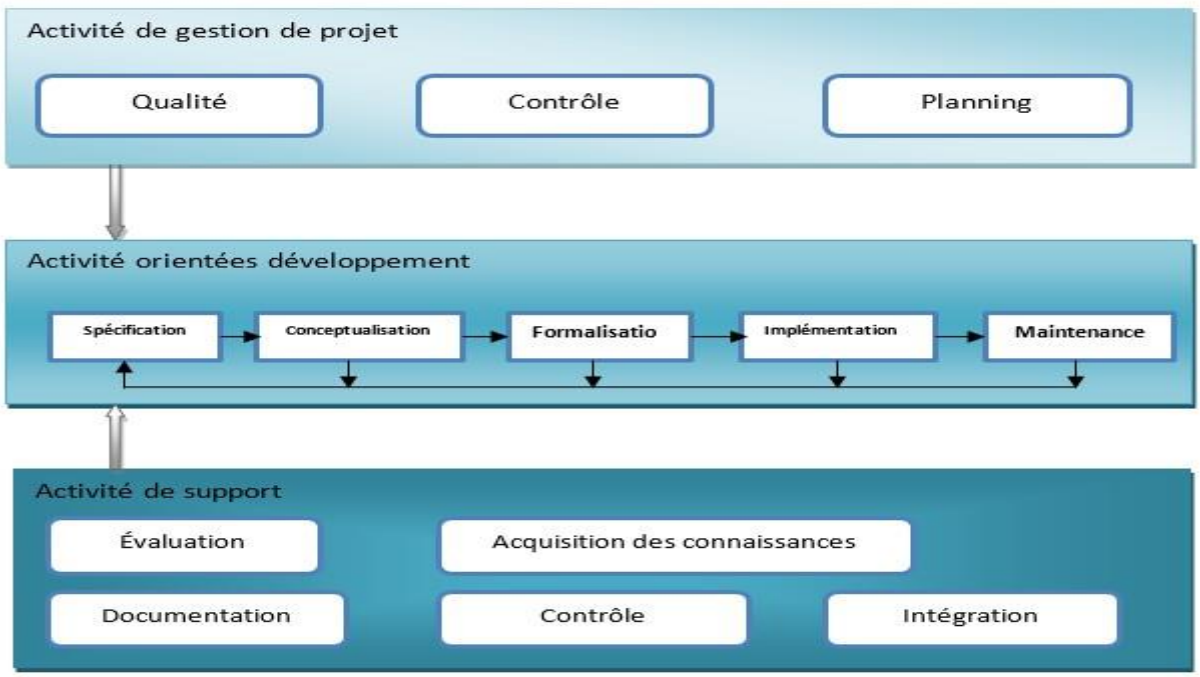


Figure 2.4 processus de développement et cycle de vie de METHONTOLOGY

2.6 Formalismes de représentation des connaissances

2.6.1 Frames

GRUBER a initialement proposé Le modèle des Frames **[M. Kifer, G. Lausen and J. Wu, 1995]** comme langage de représentation d'ontologies, Le principe de ce modèle est de décomposer les connaissances en classes (ou frames) qui représentent les concepts du domaine. A un frame est rattaché un certain nombre` d'attributs (slots), chaque attribut pouvant prendre ses valeurs parmi un ensemble de facettes (facets). Une autre façon de présenter ces attributs est de les considérer comme des relations binaires entre classes dont le premier argument est appelé domaine (domain) et la deuxième portée (range). Des instances des classes, correspondant à l'extension de chaque concept, peuvent être ajoutées, ainsi que des fonctions qui sont des types particuliers de relations liant un ensemble de classes à une valeur calculée à partir des valeurs des attributs des classes. La spécification de propriétés conceptuelles des attributs (ou relations) recourt a` des formules de la logique du premier ordre. **[F. Amourache, 2008]**

2.6.2 Logiques de descriptions

Les logiques de description (LDs) découlent directement des travaux fondateurs de Bachmann et de son système KL-ONE. Depuis le début des années 90, la recherche en logique de description s'est considérablement développée. Les logiques de description peuvent être considérées comme un fragment de la logique du premier ordre, dans lequel les formules ont une variable libre pour les descriptions de concept et deux variables libres pour les descriptions de relations. **[F. Amourache, 2008]**

Une LD est composée de deux parties :

- Un langage terminologique TBOX : définition des notions basiques ou dérivées et de comment elles sont reliées entre elles. Ces informations sont "génériques" ou "globales", vraies dans tous les modèles et pour tous les individus. La première colonne du tableau présente un exemple de T-Box, Les prochains paragraphes explicitent divers aspects des T-Box en se référant à cet exemple.

T-Box	A-Box
Femelle $\sqsubseteq \top$ Male $\sqsubseteq \top$ Femelle $\sqsubseteq \top \sqcap \neg$ Male Male $\sqsubseteq \top \neg$ Femelle Animal $\equiv$ Male $\sqcup$ Femelle Humain $\sqsubseteq$ Animal Femme $\equiv$ Humain $\sqcap$Femelle Homme $\equiv$ Humain $\sqcup \neg$ Femelle Mère $\equiv$ Femme $\sqcup$ $\exists$ relationParentEnfant Père $\equiv$ Homme $\sqcup$ $\exists$ relationParentEnfant MèreSansFille $\equiv$ Mère $\sqcap \forall$relationParentEnfant.$\neg$ Femme relationParentEnfant $\sqsubseteq \top_R$	Humain (Amel) Femelle(Amel) Femme (Lynda) Humain (Sofiane) $\neg$ Femelle(Sofiane) Homme (Amar) relationParentEnfant (Lynda, Amel) relationParentEnfant (Sofiane, Amar)

Table 2.1 Une base de connaissances composée d'une T-Box et d'une A-Box

- Les entités atomiques : Les concepts atomiques et les rôles atomiques constituent les entités élémentaires d'une T-Box. Les noms débutant par une lettre majuscule désignent les concepts, alors que ceux débutant par une lettre minuscule

dénomment les rôles (par exemple : les concepts : Femelle, Mâle, Homme et Femme, et le rôle relation Parent Enfant.

- Les concepts et les rôles atomiques prédéfinis : Les LDs prédéfinissent minimalement quatre concepts atomiques : le concept τ et le rôle $^{\wedge}$R, les plus généraux de leur catégorie respective, et le concept $^{\wedge}$ainsi que le rôle $^{\wedge}$R les plus spécifiques (c'est-à-dire l'ensemble vide).
- Les entités composées : Les concepts et les rôles atomiques peuvent être combinés au moyen de constructeurs pour former respectivement des concepts et des rôles composés. Par exemple, le concept composé Mâle ∩ Femelle résulte de l'application du constructeur ∩ aux concepts atomiques Mâle et Femelle. Le concept Mâle ∩ Femelle s'interprète comme l'ensemble des individus qui appartiennent aux concepts Mâle et Femelle. Les différentes LDs se distinguent par les constructeurs qu'elles proposent. Plus les LDs sont expressives, plus les chances sont grandes que les problèmes d'inférence soient non décidables ou de complexité très élevée. Par contre, les LDs trop peu expressives démontrent une inaptitude à représenter des domaines complexes.
    - **La définition formelle de T-Box :** Une T-Box contient des axiomes terminologiques de la forme C Ξ D ou C ⊆D. La première sert à énoncer des relations d'équivalence entre concepts, alors que la seconde permet d'exprimer des relations d'inclusion. Une interprétation I satisfait un axiome C ⊆D si et seulement si CI ´I DI. Une interprétation I satisfait un axiome C Ξ D si et seulement si CI = DI. Une interprétation satisfait une T-Box (est un modèle de T-Box) si et seulement si l'interprétation satisfait tous les axiomes de la T-Box.
    - **un langage assertion el ABOX :** défini par un ensemble d'individus désignant des objets nécessairement différents dans toute interprétation et d'instances de concepts et de relations vérifiées par ces individus. Une ABox contient un ensemble d'assertions sur les individus : (1) **des assertions d'appartenance** et (2) **des assertions de rôle**.

  Chaque ABox doit être associé à une T-Box, car les assertions s'expriment en termes de concepts et de rôles de la T-Box. La deuxième colonne du Tableau 1 illustre un exemple d'A-Box. Une A-Box d´désigne des individus dans ses

assertions par des noms qu'elle leur donne. L'exemple du Tableau 2.1 comprend les individus nommés suivants : Amel, Sofiane, Amar et Lynda.

Dans ce qui suit nous représentons par les lettres à, b les individus nommés. Une fonction d'interprétation assigne à chacun de ces noms à, un individu aï tel qu'aï ˆIDI. Les moteurs d'inférence pour les logiques de descriptions adoptent généralement l'hypothèse de noms uniques, c'est-à-dire que pour tout individu nommé $aetb \Rightarrow aI^{1}bI$.

Chaque assertion d'appartenance d'une ABox (notée C(a)ou à : C), déclare que pour cette A-Box, il existe un individu nommé a, membre du concept C de la T-Box associée. Une interprétation satisfait une assertion d'appartenance C(a) si et seulement si aï ICI. Une assertion de rôle, de la forme R(a, b)(ou (a, b) : R) indique que pour cette ABox, il existe un individu nommé a qui est en relation avec un individu nommé b par le rôle R (défini dans la T-Box associée), tel que a fait partie du domaine de R et b fait partie de l'image (le Co-domaine) de R. Une interprétation satisfait une assertion de rôle R (a, b) si et seulement si (aï, bi) ˆI RI.

2.6.3 Réseaux sémantique

Les réseaux sémantiques ont été proposés par Quillian en 1968 pour représenter explicitement un modèle psychologique de la mémoire associative humaine.

Un réseau sémantique est composé d'une part, de nœuds qui peuvent représenter indifféremment des objets, des concepts ou des évènements et d'autre part, de liens (arcs orientés étiquetés) entre les nœuds qui représentent leurs relations. Chaque réseau sémantique est donc une structure de graphe dédiée à la description d'objets et de leurs relations binaires.

De plus, ils peuvent aussi servir de représentation graphique de prédicats binaires de la logique des prédicats. Les symboles de variables, fonctions et prédicats sont alors représentés par un réseau de nœuds : les variables et les fonctions sont représentés par des nœuds et les prédicats binaires par des liens. **[43]**

2.7 Les langages de représentation d'ontologies

Pour le développement d'ontologies, il est nécessaire de choisir un langage ou un ensemble de langages dans lesquelles cette dernière est exprimée et utilisée, pour cela nous allons présenter les différents types de langages tel-que XML; RDF/RDFs; OWL et enfin DAML-OIL.

2.7.1 XML (Extensible Markup Language)

XML est un langage de description de documents utilisé pour décrire la structure des documents qui d´dérive de SGML (Standard Generalized Markup Language) et HTML (Hyper Text Markup Language). Il s'agit alors d'un langage formé de balises qui permet de structurer les documents ou bien qui d´écrit comment les balises sont utilisées pour diviser les documents de données structurées en différentes parties et la manière d'identifier ces parties. Il est utilisé dans la plupart des projets publication sur le web et dans les bases de données. Ce dernier est beaucoup plus utilisé pour stocker des documents que pour échanger des données. **[44]**

2.7.2 RDF (Ressource Description Framework)

Développe **[45]** par le W3C (World Wide Web Consortium), RDF est le langage de base de web sémantique dont chaque ressource est pourvue d'un identifiant URI (Uniform ressource identifier).Tout document RDF est composé d'un ensemble de triplets sujet, prédicat, objet. Un ensemble de tels triplets est appelé un graphe RDF. Ceci peut être illustré par un diagramme composé de nœuds et d'arcs orientés, dans lequel chaque triplet est représenté par un lien nœud-arc-nœud (d'où le terme de "graphe"), ou :

- **Sujet :** Représenté la ressource d´écrite, pointé par une URI
- **Prédicat :** Représente la relation utilisée pour d´écrire une ressource
- **Objet :** représente la valeur d'une propriété associée à une ressource spécifique.
- RDF peut aussi être exprimé en XML afin notamment d'´échanger ses données avec les agents logiciels du Web ou autres. A propos du langage, un document RDF est délimité par l'élément rdf : RDF qui comporte un ou plusieurs éléments rdf : Description pour chacune des descriptions de ressources comprises dans le document. Chaque

description comprend un attribut rdf : about qui pointe vers l'URI de la ressource à d´écrire et un à plusieurs éléments représentant chacun un prédicat. Lorsqu'un prédicat a pour valeur une autre ressource, l'attribut rdf : ressource pointera vers son URI. Les autres caractéristiques de RDF permettent la composition des énoncés en structures plus complexes comme le groupement de sujets et/ou d'objets en listes énumérées ou bien la réification d'énoncés, i.e. la création de nouveaux énoncés a` partir d'énoncés existants. RDF répond aux besoins de la plupart des outils d'annotation. En effet, les documents RDF sont des documents XML valides, leur modélisation sous forme de réseau sémantique apporte une flexibilité nécessaire et il est possible de réutiliser des énoncés existants pour composer des documents RDF plus complexes.

Exemple : Amayas a 19 ans et habite à Tizi-Ouzou

```
<rdf :RDF>
<rdf :Description about='Amayas'>
<rdf :Porperty about='ville'>
    Tizi-Ouzou
</rdf :Porperty>
<rdf : Porperty about='age'>
    19
</rdf :Porperty>
</rdf :Description>
</rdf :RDF>
```

Figure 2.5 Exemple de document RDF

Par contre, RDF ne fournit pas de mécanisme de contrainte de classes ou de types pour les différentes parties du triplet. Il n'est donc pas assez puissant pour représenter de vraies ontologies avec un système de raisonnement approprié.

2.7.3 RDFS (RDF Schéma)

RDFS décrit les ressources avec des classes, les propriétés et les valeurs, autrement dit c'est un langage extensible de représentation des connaissances et qui

permet aussi de définir le vocabulaire pour décrire des classes et des propriétés hiérarchisées en taxinomies. **[46]**

Sur l'exemple d'Amayas, nous définissons le concept de personne, une taxinomie de concepts et l'instance Amayas.

2.7.4 OWL (Web Ontology Language)

Le langage d'ontologie Web OWL est conçu pour être utilisé par les applications qui ont besoin de traiter le contenu de l'information au lieu de simplement présenter des informations pour les humains. OWL est une extension de RDF et RDFS qui fournit un vocabulaire supplémentaire avec une sémantique formelle. OWL a trois sous-langages plus expressifs : OWL Lite, OWL DL et OWL Full. **[47].**

```
<rdf :RDF>
<rdfs :Class rdf :about='Personne'>
  <rdfs :subClassOf rdf :resource='Thing'/>
</rdfs :Class>

<rdf :Property about='age'>
<rdfs :domain rdf :resource='Personne'/>
<rdfs :range rdf  :resource='xsd :integer'/>
</rdf :Property>
<rdf :Property about='ville'>
<rdfs :domain rdf :resource='Personne'/>
<rdfs :range rdf :resource='xsd :string'/>
</rdf :Property>
</rdf :RDF>

    <Personne rdf :ID='Amayas'>
    <age rdf :resource='19' />
    <ville rdf :resource='Tizi-Ouzou' />
    </Personne>
```

Figure 2.6 – Exemple de document RDFs

- **OWL Lite**

Moins Complexe, destiné aux utilisateurs qui ont besoin d'une hiérarchie de concepts simple.

- **OWL DL**

Complexe, permet une expressivité plus importante et garantit la complétude des raisonnements.

- **OWL FULL**

Plus complexe, permet le plus haut niveau d'expressivité.

- **DAML-OIL :**

DAML + OIL est un langage de balisage sémantique pour les ressources web. Il s'appuie sur les standards du W3C antérieurs, tels que RDF et RDF schéma, et s'étend ces langues avec riches primitives de modélisation. DAML + OIL fournit des primitives de modélisation couramment trouvés dans les langues basé sur les images. Il a été construit a` partir de la langue originale de l'ontologie DAML-ONT

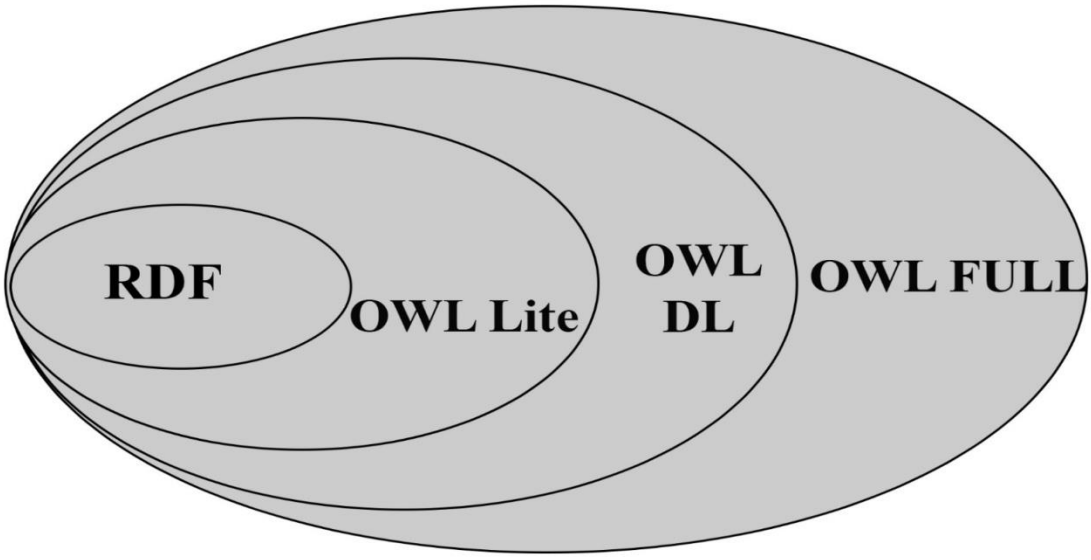


Figure 2.7 Hiérarchie de langage OWL

Dans le but de combiner de nombreux composants linguistiques. Le langage a une sémantique propre et bien défini. **[48]**

2.8 Les outils d'édition

Il existe plusieurs éditeurs d'ontologies qui utilisent des formalismes variés et offrent différentes fonctionnalités. Dans ce qui suit nous allons définir quelques outils : **Onto lingua**, **Web Onto**, **Protégé**.

2.8.1 Ontolingua

[Farquhar & Fikes, 1996] Développer à l'université de Stanford, l'architecture de son serveur permet d'accéder à une bibliothèque d'ontologies et à des traducteurs de langages de programmation (Prolog, Lomm) et possède un éditeur de création et de parcours des ontologies. Il permet trois types d'interaction les collaborateurs qui souhaitent écrire et examiner des ontologies à distance, Applications éloignées susceptibles de vouloir interroger et modifier des ontologies sur le serveur via l'Internet et les applications locales.

Pour intégrer les ontologies Ontolingua on distingue trois possibilités

Inclusion : en utilisant et incluent les définitions d'autres ontologies.

Restriction : en important les définitions depuis d'autres ontologies et les rendre plus spécifiques.

Raffinement polymorphe : on redéfinit une définition importée depuis n'importe quelle ontologie.

2.8.2 Web Onto

Developer au Knowledge Media Institute à L'Open University. C'est un outil accessible sur Internet et principalement graphique permettant de construire comparativement des ontologies. Il permet une visualisation graphique et séparée des différents composants d'une ontologie (classes, instances, relation, règles, procédures) adaptée à la construction d'ontologies de grande taille. Et il offre éventuellement des services différentiels, basés sur le langage OCML permettant de répondre à des requêtes et des vérifications de cohérence. **[49]**

2.8.3 Protégé

Protégé **[Noy, et al. 2000]** C'est un éditeur d'ontologie qui a été développé par le Département d'Informatique Médicale de l'Université de Stanford; PROTEGE est une plate-forme Open Source autonome, qui fournit un environnement graphique

permettant l'édition, la visualisation et la vérification des contraintes (contrôle) de l'ontologie.

Le modèle de représentation de connaissances de PROTEGE, est issu du modèle des frames. Ce dernier contient des classes (pour modéliser les concepts), des slots (pour modéliser les attributs des concepts) et des facettes (pour définir les valeurs des propriétés et des contraintes sur ces valeurs), ainsi que des instances des classes.

L'édition des listes de ces trois types d'objets se fait par l'intermédiaire de l'interface graphique, sans avoir besoin d'exprimer ce que l'on a à spécifier dans un langage formel : il suffit juste de remplir les différents formulaires correspondant à ce que l'on veut spécifier. L'interface très bien conçue, et l'architecture logicielle permettant l'insertion de plugins pouvant apporter de nouvelles fonctionnalités, notamment les plugins pour gérer les représentations sous forme graphique, par exemple OWLViz et la prise en charge de nouveaux langages ,ont participé à son succès et le rendent l'éditeur d'ontologie jouissant de la plus grande renommée a` l'heure actuelle, servant de référence pour une importante communauté d'utilisateurs.

2.9 Domaines d'application d'ontologies

2.9.1 Les ontologies et la représentation des connaissances

Les ontologies sont apparues en informatique, plus précisément en Ingénierie des Connaissances, dans le cadre des démarches d'acquisition des connaissances pour les systèmes à base de connaissances (SBC). Les SBC proposaient alors de spécifier, d'un côté des connaissances du domaine modélisé, et de l'autre, des connaissances de raisonnement qui manipule et utilise ces connaissances du domaine. L'idée de cette séparation modulaire était de construire mieux et plus rapidement des SBC en réutilisant le plus possible des composants génériques, que ce soit au niveau du raisonnement ou des connaissances du domaine.

En conclusion, l'objectif de l'ingénierie ontologique est de diversifier les applications des Systèmes à Base de Connaissances (SBC), et de permettre une représentation des connaissances indépendantes de ces diverses applications, de manière a` assurer sa portabilité d'une application a` l'autre. **[Bachimont 2003]**

2.9.2 Les ontologies et le web sémantique

Le Web actuel est essentiellement syntaxique, la structure des ressources étant bien définie, mais leur contenu restant inaccessible aux traitements machines, seuls les humains étant capables de l'interpréter. **[Lee 2002]**

Le Web sémantique a alors l'ambition de lever cette difficulté en associant aux ressources du Web des entités ontologiques comme références sémantiques, ce qui permettra aux différents agents logiciels d'accéder et d'exploiter directement le contenu des ressources et de raisonner dessus. Ce référencement sémantique peut aussi résoudre les problèmes d'interprétation des ressources informationnelles provenant des applications hétérogènes et réparties et de permettre ainsi à ces applications d'être intégrées sémantiquement. **[Uschold 2002]**

L'architecture du Web sémantique repose sur une hiérarchie des langages d'assertion et de description d'ontologies ainsi que sur un ensemble de services pour l'accès aux ressources au moyen de leurs références sémantiques, pour gérer l'évolution des ontologies, pour l'utilisation des moteurs d'inférences capables d'effectuer des raisonnements complexes ainsi que des services pour la vérification de la validité sémantique de ces raisonnements. **[Oberle 04].**

2.10 Conclusion

Nous avons détaillé dans ce chapitre la notion d'ontologie, en essayant d'éclaircir la notion en présentant certaines définitions. Nous avons montré aussi les différents intérêts des ontologies et sa structure, ainsi que les différentes classifications, les méthodologies les plus représentatives de leur construction. Ensuite nous avons présenté les principaux formalismes de représentation de connaissances. Finalement nous avons étudié les langages d'ontologies et les outils nécessaires aux développements des ontologies dans divers domaines d'applications. La construction des ontologies est un processus difficile et fastidieux. Les outils que nous avons présentés dans ce chapitre contribuent à amortir cet effort à travers les facilités et les multiples avantages qui portent les langages de développement d'ontologies dans de nombreux usages.

Au cours du chapitre suivant, nous nous intéressons aux étapes de construction de

notre ontologie de domaine médicale.

Chapitre 3

Conception de l'ontologie

Introduction

Ce chapitre présente notre contribution à la problématique posée dans ce mémoire, à savoir la construction d'une ontologie de domaine médicale. Cette ontologie est appelée une ontologie de domaine, d'abord en suivant un processus de construction d'une ontologie de domaine partant de connaissances brutes et arrivant à une ontologie opérationnelle. Par la suite, nous décrirons comment nous avons utilisé l'outil PROTEGE pour éditer notre ontologie en langage OWL

3.1 Processus de construction d'une ontologie de domaine

Nous avons suivi un processus de construction d'une ontologie de domaine partant de connaissances brutes et arrivant à une ontologie de domaine opérationnelle représentée par langage OWL. Ce processus est inspiré de la méthodologie METHONDOLOGIE qu'on a bien détaillé dans le premier chapitre, et il est composé de cinq étapes :

- Evaluation des besoins.
- Conceptualisation.
- Formalisation.
- Implémentation.
- Vérification et Evaluation.

3.2 Construction d'une ontologie de domaine

Pour réaliser cette ontologie de domaine nous suivons le processus décrit précédemment :

3.2.1 ´Evaluation des besoins

Pour réaliser cette ontologie de domaine nous suivons le processus décrit précédemment :

- **Le domaine de connaissance** : Le domaine médical.
- **L'objectif :** Partager de façon collaborative les connaissances médicales et, Faciliter la recherche et l'intégration d'informations provenant des multiples Sources d'information médicales.
- **Les utilisateurs:** Les utilisateurs de cette ontologie sont : les *médecins*
- **Les sources d'informations :** L'encyclopédie médicale, interviews avec les Médecins, sites web médicaux (Doctissimo, Eureka sente, experts du domaine...)
- **La portée de l'ontologie** : Médecin, Médicament, maladie, diagnostiqué ...

3.2.2 Conceptualisation

Dans cette étape on distingue les principales tâches suivantes :

Construire le glossaire de termes

Construire un glossaire de termes est la première tâche à effectuer dans l'étape de conceptualisation. Il recueille et décrit tous les termes (concepts, instances, attributs, relations entre les concepts, etc.) qui sont utiles et potentiellement utilisables dans le domaine que nous allons représenter leurs descriptions détaillées et non ambiguës en langage naturel.

Le tableau suivant présente la liste de quelques termes utiliser dans notre ontologie :

Terme	Signification
Maladie	La maladie est une altération des fonctions ou de la santé d'un organisme, se traduisant par des signes .elle concerne un ou plusieurs patients.
Médecin spécialiste	Il est un professionnel de santé. Il peut exercer à l'hôpital ou avoir une activité libérale.
Médicament	La maladie est une altération des fonctions ou de la santé d'un organisme, se traduisant par des signes .elle concerne un ou plusieurs patients.
Donne comme effet	Il est un professionnel de santé. Il peut exercer à l'hôpital ou avoir une activité libérale.
Symptôme	Un médicament est substance présentée comme possédant des propriétés curatives, préventives ou administrée en vue d'établir un diagnostic .un médicament est le plus souvent destiné à guérir à favoriser la guérison , à soulager ou à prévenir des maladies humaines ou animales .
Diagnostique	Relation entre médicaments et effet secondaire.

Maladies endocriniennes	Manifestation d'une maladie observée par le médecin lors de l'examen de son patient.
Maladie dermatologique	Le raisonnement menant à l'identification de la cause (l'origine)d'une défaillance, d'un problème ou d'une maladie, ou tout simplement à la détermination d'une espèce biologique par rapport à une autre (taxinomie) ,à partir des caractères ou symptômes relevés par des observations ,des contrôles ou des tests .

Table 3.1 Glossaire de quelques termes

- Construction de diagramme des relations binaires et des attributs

Une relation binaire permet de relier deux concepts entre eux (un concept source et un concept cible). _ Si R est une relation entre deux concepts C1 et C2 alors pour tout couple d'instances des concepts C1 et C2, il existe une relation de type R qui lie deux instances de C1 et C2 _. Cette tâche permet de représenter d'une manière graphique les différentes relations qui existent entre les divers concepts que ce soit de même ou de différentes hiérarchies.

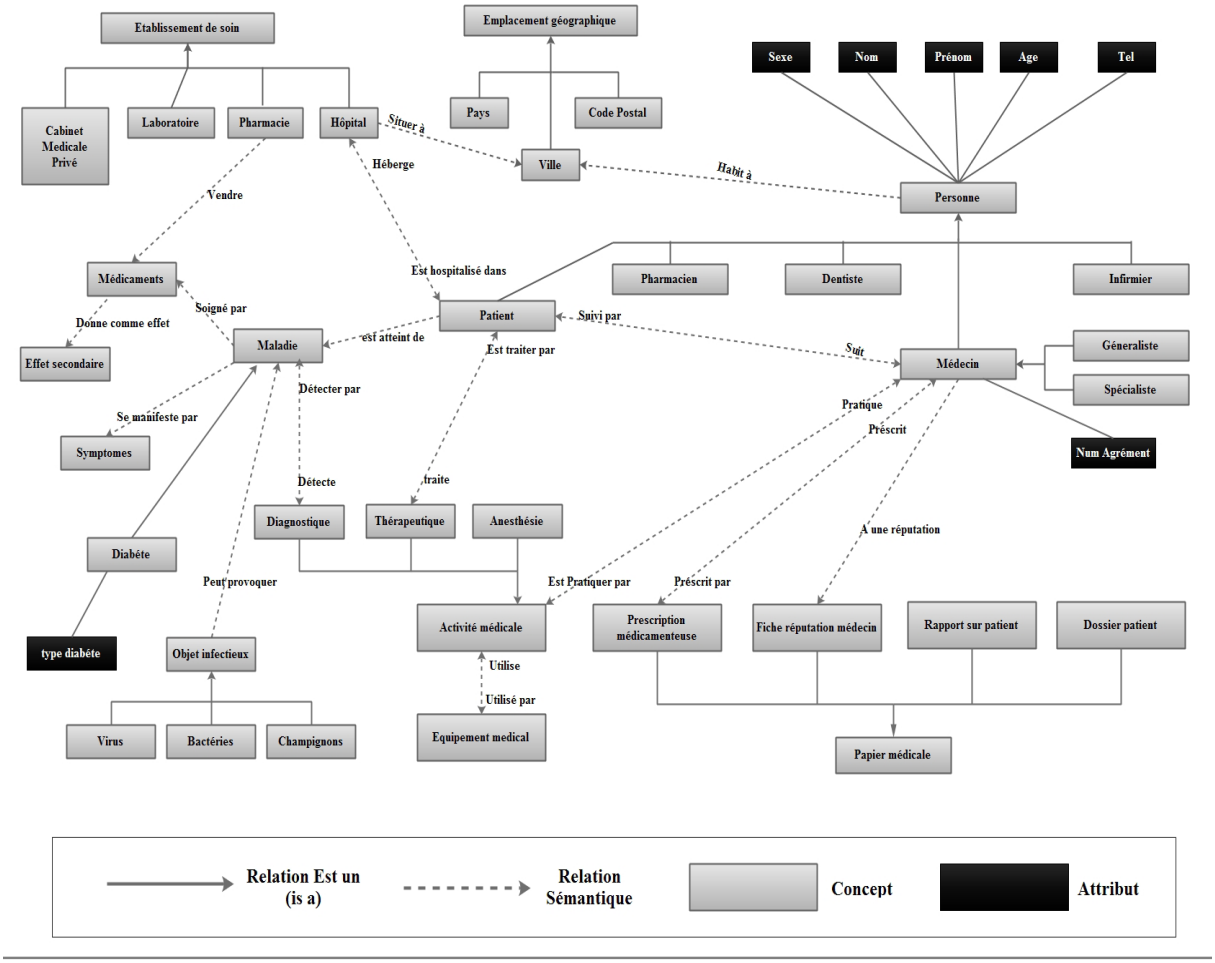


Figure 3.1 Diagramme des relations binaires

- **Construction d'un dictionnaire de concepts**

Une fois la taxonomie de concepts et le diagramme de relations binaires sont

Effectués, il faut donc spécifier les propriétés qui décrivent chaque concept de la hiérarchie dans un dictionnaire de concepts. Un dictionnaire de concepts contient tous les concepts du domaine, leurs synonymes, leurs acronymes, leurs attributs et leurs relations. Les relations spécifiées pour chaque concept sont celle où leur domaine est le concept lui-même.

Ce tableau représente le dictionnaire de quelques concepts utilisés dans notre Ontologie.

Nom du concept	Attributs	Relations	Acronyme	Synonymes

Maladie	/	Est_diagnostique_par Est_traitée_par Présente	Mal	Souffrance
Médecin Spécialiste	Numéro Agrément Nom Prénom Age Téléphone Sexe	Traite	Dr	Docteur
Médicament	/	Donne_comme_effet	Médoc	Traitement

Table 3.2 dictionnaire de quelques concepts

- **Déterminer la liste des relations binaires**

Le but de cette tâche consiste à construire une table de relations binaires

Décrites en détail. Pour chaque relation utilisée dans le diagramme des relations Binaires, nous définissons le nom de la relation, le nom des concepts sources et

Cibles, le nom de la relation inverse et les cardinalités source et cible. Le tableau

Suivant présente quelques relations binaires utilisées dans notre ontologie.

Nom de relation	Concept source	Concept cible	Relation inverse	Cardinalité
Traite	Médecin spécialiste	Maladies	/	1. n
Présente	Maladies	Symptômes	/	1. n
Est_diagnostique_par	Maladies	Diagnostique	/	1. n
Est_traitée_par	Maladies	Traitement	/	1. n

Donne_comme_effet	Médicaments	Effet secondaire	/	1. n

Table 3.3 Table de quelques relations binaire

- Construction de la table d'attributs

La table des attributs comporte une description détaillée des attributs inclus

Dans le dictionnaire de concepts, et l'ensemble de contraintes et de restrictions sur ces valeurs. Le tableau suivant présente quelques attributs utilisés dans notre ontologie.

Nom attribut	Type	Card (min/max)	Valeur par défaut	Domaine de valeur
Nom	String	1..1	-	-
N° d'agrément	String	1..1	-	-
Age	Integer	1..1	-	-
Type médicament	String	1..6	-	Comprimé, gélule, sirop, crème, injectable

Table 3.4 Table des attributs

- Construction de la table des axiomes

Dans cette étape, nous définissons les concepts au moyen d'expressions logiques possible. Pour chaque axiome, il faut spécifier la description de l'axiome en langage naturel, l'expression logique qui décrit formellement l'axiome en logique du premier ordre, les concepts, les relations et les variables utilisées. Le tableau suivant présente quelques axiomes utilisés dans notre ontologie.

Nom de concept	Description de l'axiome en langage naturel	expression logique
Ophtalmologue	Un Ophtalmologue traite les maladies de l'œil	∀x Ophtalmologue(X) ∧ ∃ Y maladie de l'œil (Y) ∧ Traite(X,Y)
Medecin spécialiste	Chaque médecin spécialiste est soit un pneumologue, cardiologue, un pédiatre	∀ X Medecin spécialiste (X) ⊃ Pneumologue(X) ∨ Cardiologue(X) ∨ Pédiatre(X)
Maladies	Chaque maladie est détectée par un ou plusieurs diagnostiques	∀ X maladie (X) ∃ Y diagnostique(Y) ∧ détecter par (x, y)
Traitement	Un Traitement est soit un Médicament, Thérapeutique, Chirurgie	∀ X Traitement (X) ⊃ Medicament(X) ∨ Thérapeutique(X) ∨ Chirurgie(X)

Table 3.5 Table des axiomes

- **Construction de la table des instances**

La table des instances décrit les instances connues ; qui sont déjà identifiées

Dans le dictionnaire de concepts. Pour chaque instance, il faut spécifier le nom de l'instance, le nom du concept où elle appartienne, ses attributs et les valeurs qui lui y sont associés. Le tableau ci-après illustre quelques instances créées.

Concept	Instance
Pneumologue	Dr Hameg
Diagnostique	Electrocardiogramme
Médicament	Antalgique
Symptôme	Constipation
Effet secondaire	Fièvre
Antidiabétiques	Insuline
Maladies endocrinienne	Goitre

Table 3.6 Table des instances

3.2.3 Formalisation

Comme cité auparavant, dans cette étape, nous utilisons le formalisme de la

Logique de description pour formaliser le modèle conceptuel que nous avons obtenu dans l'étape de conceptualisation. Le résultat est une base de connaissances en logique de description composé de deux parties T-BOX et A-BOX.

- **Le niveau terminologique ou T-Box**

Voici quelques définitions et axiomes terminologiques représentées dans le tableau suivant :

Axiome terminologique
Medecin Spécialiste ⊆Thing
Effet Secondaire ⊆Thing
Pneumologue ⊆ Medecin Specialiste
Traitement ⊆ Thing
Medicament ⊆ Traitement
Antibiotique ⊆ Medicament ⊆ Traitement
Diagnostique ⊆ Thing ∩ (∃détecte. Maladie)
Endocrinologue ⊆ Medecin Specialiste ∩ (∃ NumAgrément.String)∩ (∃nom.String)∩ (∃prenom.String) ∩ (∃ Telephone.String)∩ (∃traite. Maladies Endocriniennes)
Maladies œil ⊆ maladie∩ (∃diagnostiquer par .Ophtalmologue)
(Pneumologue∩ Cardiologue ∩ Viscéralgie ∩ pédiatre) ⊆ Medecin Specialiste

Table 3.7 Axiomes terminologiques (T-Box)

Construction d’ABox

Voici quelques assertions sur les individus représentées dans le tableau suivant:

Assertion sur les individus
• Pneumologue (Dr_Hameg) • Diagnostique par (maladies de l’œil, ophtalmologue) • Soigné par (insuline) • Viséraliste (Dr_Hamrioui) • Cardiologue (Dr_Hemdani) • Médicament (antalgique (panadole)) • Médicament (antidiabétiques (Oraux (amarel)) • Thérapeutique (chimiothérapie • Diagnostique (biopsie) • Traitement (chirurgie) • Douleur (arthralgie) • Maladies (maladies dermatologique (acné) • Symptôme (fatigue)

Table 3.8 Assertions sur les individus (ABox)

Chapitre 4

Implémentation et Réalisation

Introduction

Dans ce chapitre nous allons présenter le travail d'implémentation qu'on a fait, qui consiste premièrement à l'édition de notre ontologie sous le langage OWL et deuxièmement le site.

4.1 Etude de protégé

Protégé est un éditeur d'ontologies hautement extensible, capable de manipuler des formats très divers, c'est aussi une librairie JAVA qui peut être étendue afin de créer des applications à base de connaissance en utilisant un moteur d'inférence pour raisonner et déduire de nouveaux faits. Incluant des plugins pour les langages

RDF, DAML+OIL et OWL pour la manipulation d'ontologies dans différents formats, il aide ou permet de construire des ontologies pour le web sémantique.

Protégé permet la création et l'édition des ontologies grâce à ces deux outils distincts :

- Protégé-Frame : Permet de créer facilement une interface graphique afin de bien gérer une ontologie, les formulaires se génèrent automatiquement en se basant sur le schéma d'ontologie créé. Il offre également la possibilité de personnaliser l'interface selon les besoins de l'utilisateur.
- Protégé-Owl : C'est une extension de protégé qui supporte le langage OWL.

Il permet de créer des classes, propriétés, instances grâce aux nombreuses propriétés offertes par OWL. Il est aussi optimale d'interroger un raisonneur afin de contrôler l'intégrité du modèle et de créer un modèle d'inférences.

4.2 Etude de Visual studio et xampp

- Visual Studio Code est un éditeur de code simplifié qui prend en charge les opérations de développement telles que le débogage, l'exécution de tâches et le contrôle de version. Il vise à fournir uniquement les outils dont un développeur a besoin pour un cours rapide de création et de débogage de code et laisse les flux de travail plus complexes à des IDE plus premium, tels que l'IDE Visual Studio.

- Xampp est un ensemble de logiciels libres. Le nom est un acronyme venant des initiales de tous les composants de cette suite. Ce dernier réunit donc le **serveur** Web Apache, la base de données relationnelle et système d'exploitation MySQL ou MariaDB ainsi que les langages scripts Perl et PHP

4.3 Choix de langage

Notre choix a été orienté vers :

- OWL qui est le langage standard de représentation et de spécification de l'ontologie, comparé à RDFS qui est insuffisant pour codifier l'ontologie en termes de fonctionnalités sémantiques. Ces raisons suivantes illustrent notre choix d'OWL DL :
  - OWL DL permet une expressivité importante qu'OWL lite.
  - OWL DL permet d'exprimer des cardinalités multiple.
  - OWL DL offre un niveau d'expressivité suffisant tout en maintenant la complétude de calculs (toutes les inférences sont calculables) et la décidabilité (leurs calculs se fait en une durée finie).
- XML c'est un langage permettant de représenter et structurer des informations à l'aide de balise que chacun peut définir et employer comme il le veut.
- Le langage Html est un langage de balisage, c'est à dire que le texte (le contenu) est structure par des balises qui en définissent la structure (le contenant).

Il s'agit d'un langage hypertexte. Cela signifie qu'il est possible de définir des liens entre plusieurs documents ou au sein d'un même document. Les documents pointes par les liens peuvent se situer sur des machines éloignées, faisant partie d'Internet.

- PHP est un langage script, orienté vers le monde de l'Internet.

4.4 Choix de l'outil

- Pour l'édition de notre ontologie nous avons choisi l'outil Protégé parce qu'il n'impose pas de méthodologie, c'est aussi un éditeur hautement extensible, capable de manipuler des formats très divers. Le support d'OWL, comme de nombreux autres formats, est possible dans Protégé grâce à un plugin dédié.

Il a la particularité de maintenir un espace de nommage unique pour tous ces cadres. Mais il faut savoir que c'est un outil qui est sensible à la casse, ainsi personne et Personne sont deux classes différentes.

- L'un des outils le plus important de VS Code est la capacité de déboguer les applications directement sur l'éditeur sans même avoir recours aux navigateurs, grâce à un système de points d'arrêt et une console de débogage intégrée qui permet de résoudre les problèmes directement dans l'éditeur

- L'utilisation de XAMPP sert à tester les sites de vos clients ou votre propre site Web avant de le télécharger sur le serveur Web distant. Ce logiciel serveur XAMPP vous donne l'environnement approprié pour tester des projets MYSQL, PHP, Apache et Perl sur l'ordinateur local

4.5 Etapes de construction de l'ontologie et le site

A partir de là nous présentons les étapes essentielles de la construction de notre ontologie sous Protégé.

4.5.1 Lancement de Protégé

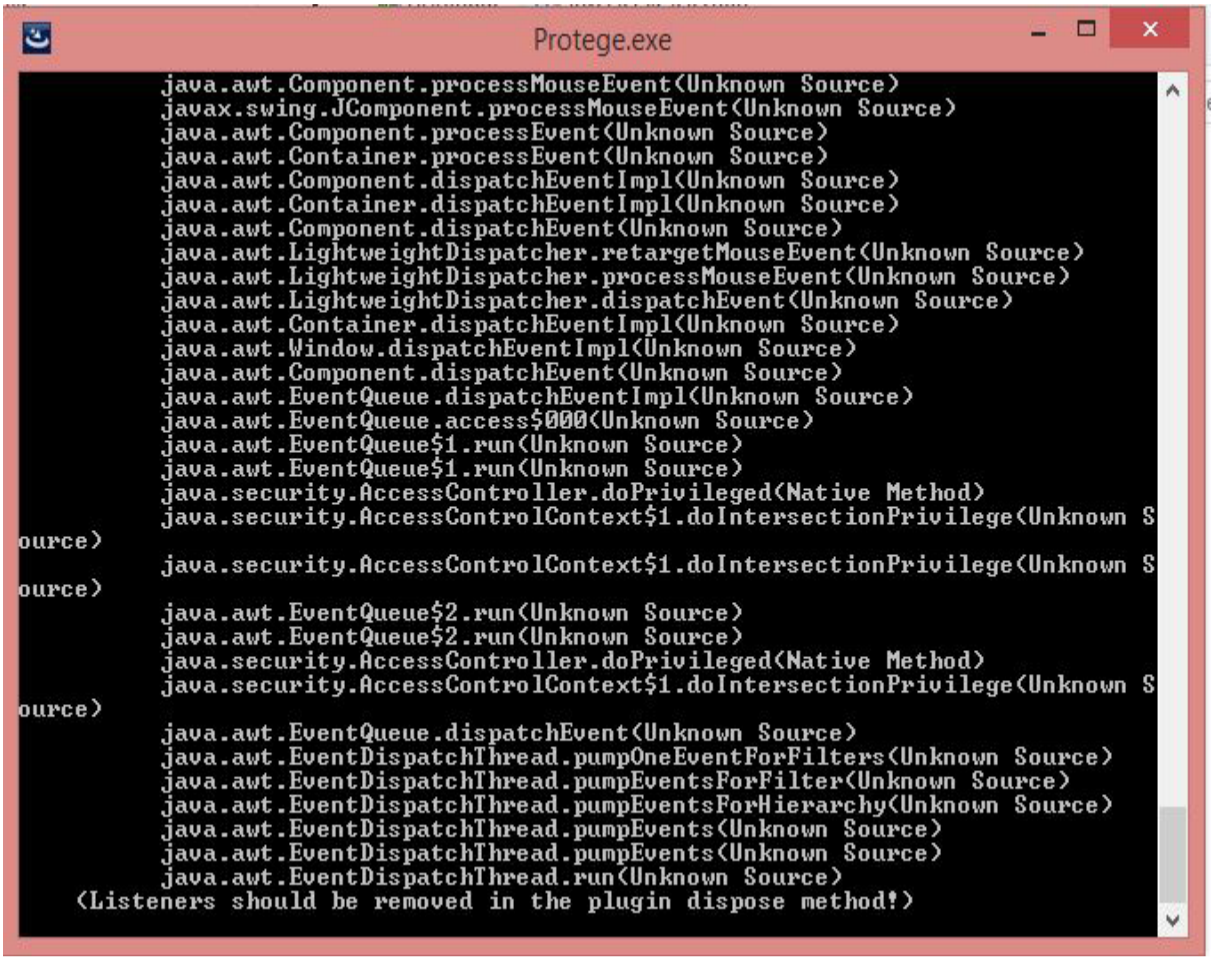


Figure 4.1 Lancement de protégé

4.5.2 Création d'un nouveau projet

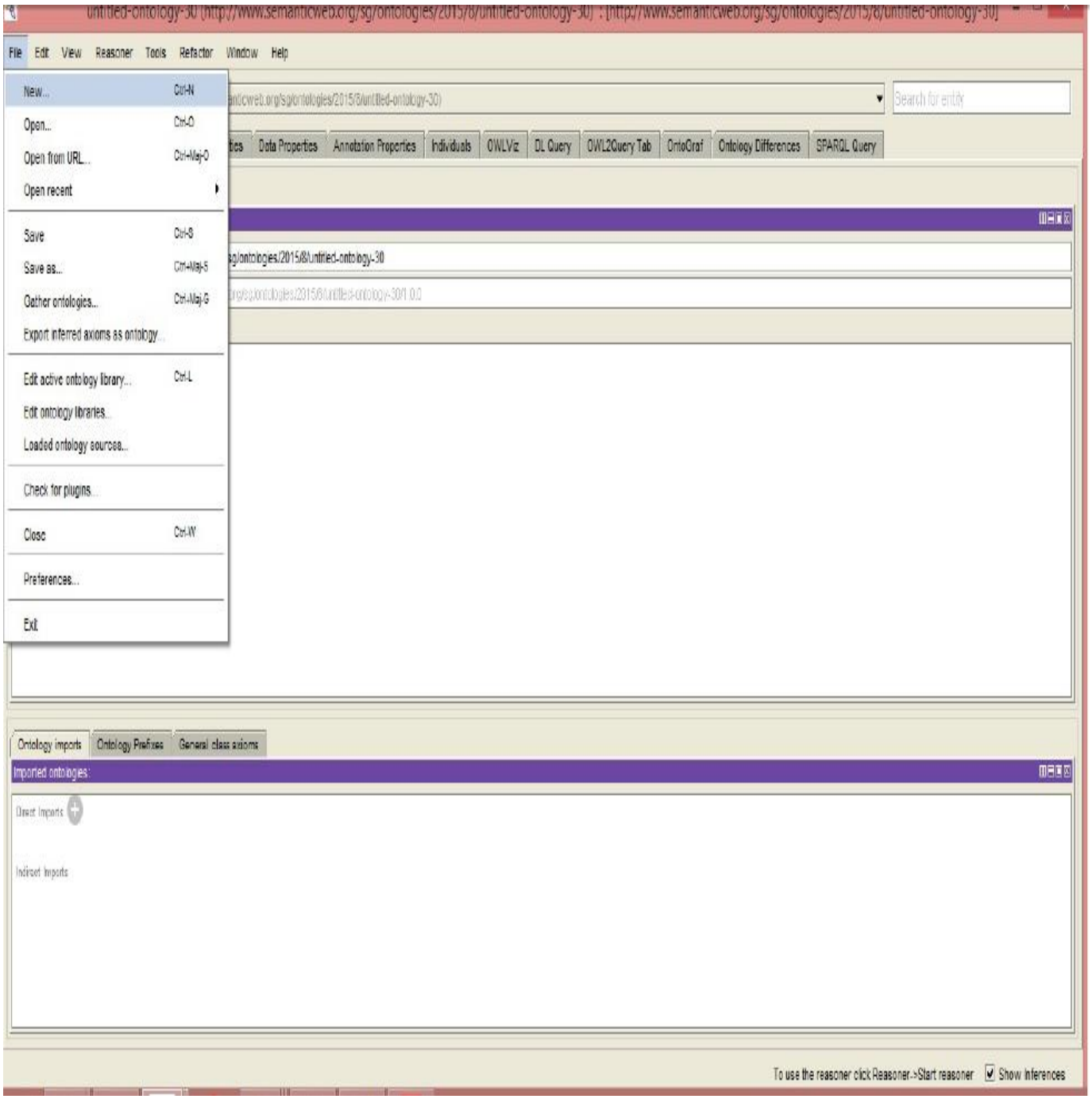


Figure 4.2 Création d'un nouveau projet

4.5.3 Création des classes : Création de la classe Maladie

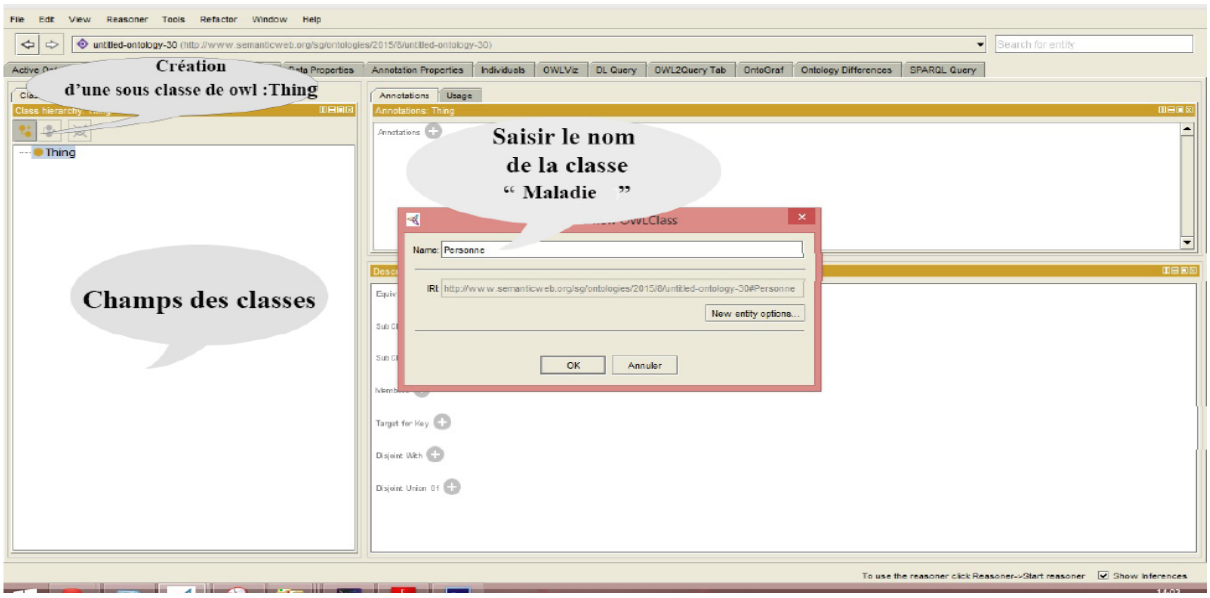


Figure 4.3 Création d'une classe

4.5.4 Création des classes disjointes

Les classes ayant le même parent sont dites disjointes, car un individu ne peut pas être une instance de plus d'une même classe.

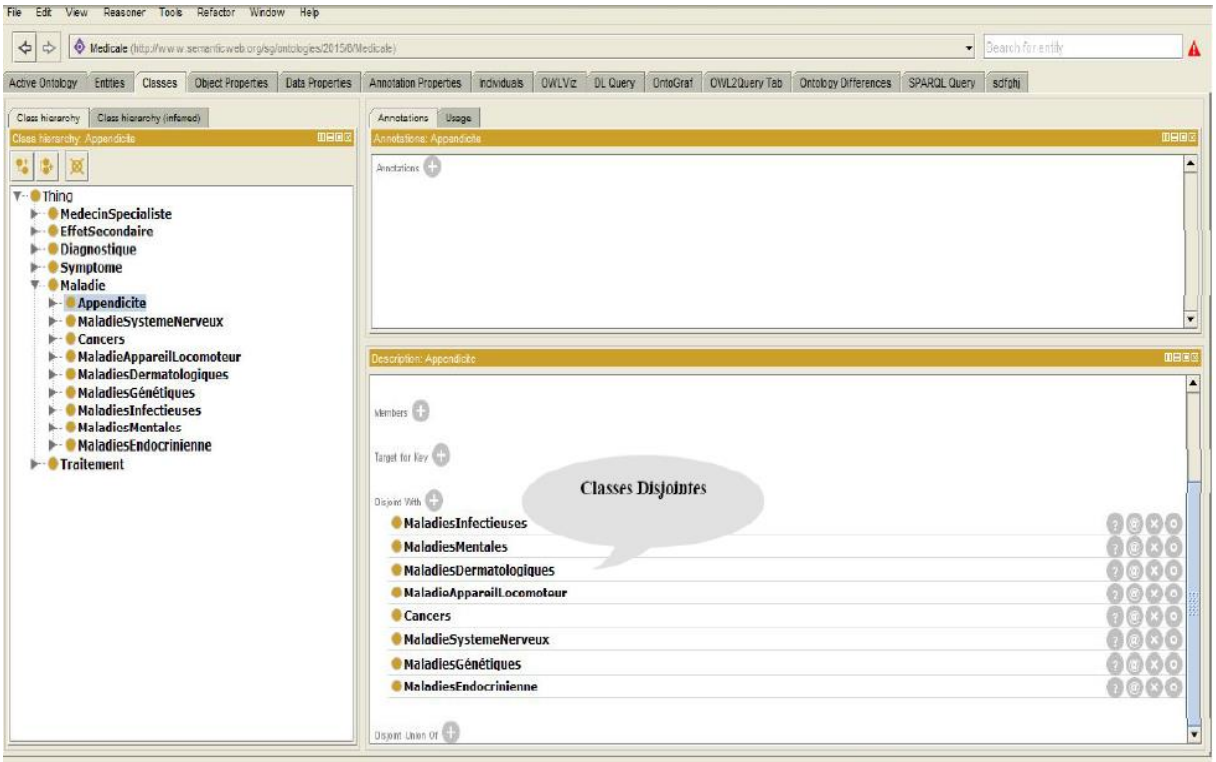


Figure 4.4 Classes disjointe

4.5.5 Création des relations

On crée la relation est ”Traite” qui a pour domaine la classe Médecin spécialiste et pour image la Maladie.

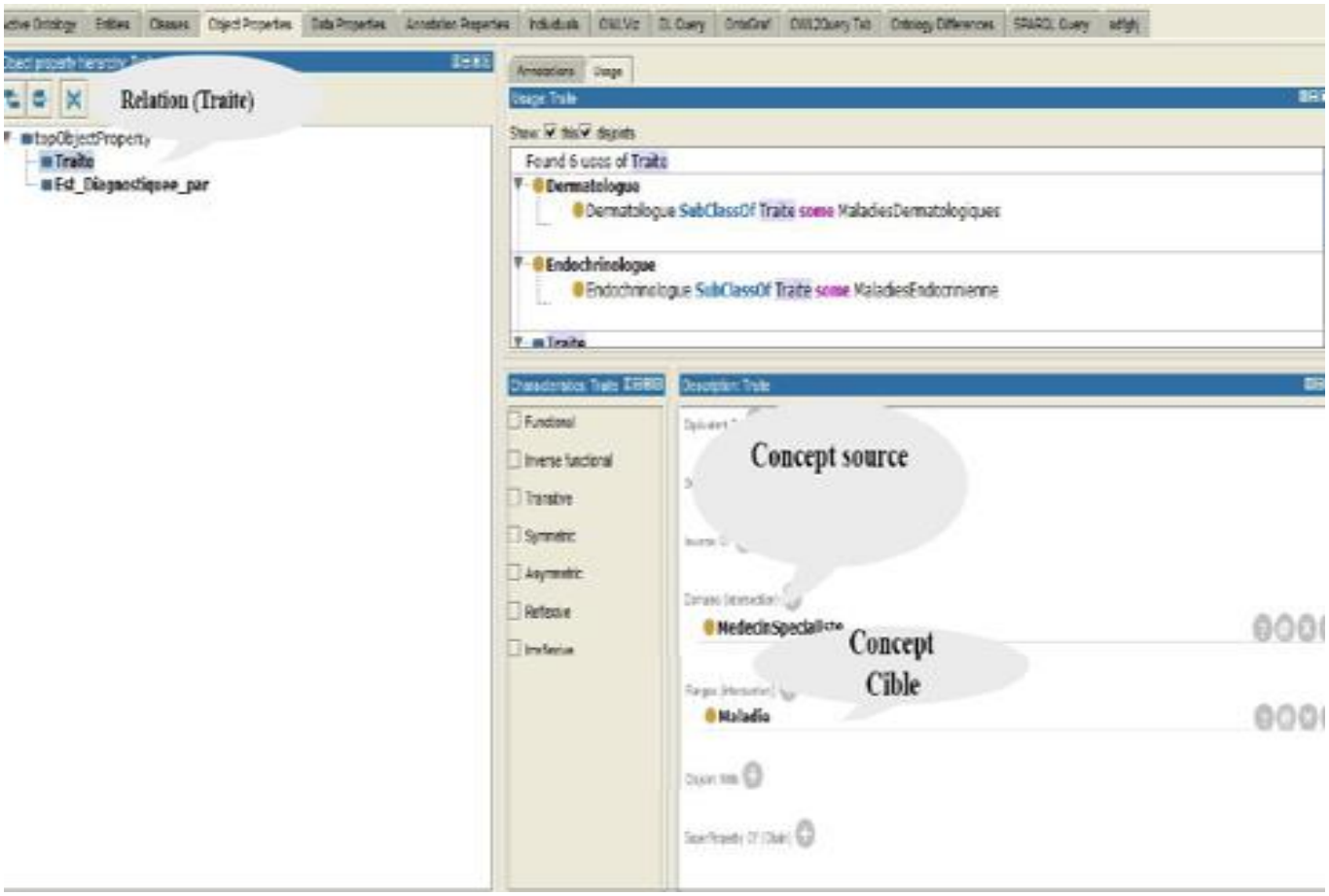


Figure 4.5 Ajout de la relation ”Traite”

4.5.6 Création des Individus

Pour chaque classe on peut créer un ou plusieurs individu on prend par exemple la classe médecin spécialiste ; on choisit Pneumologue on le renomme Dr_Hameg_ puis on remplit les champs (relations ou attributs) associer à ce médecin.

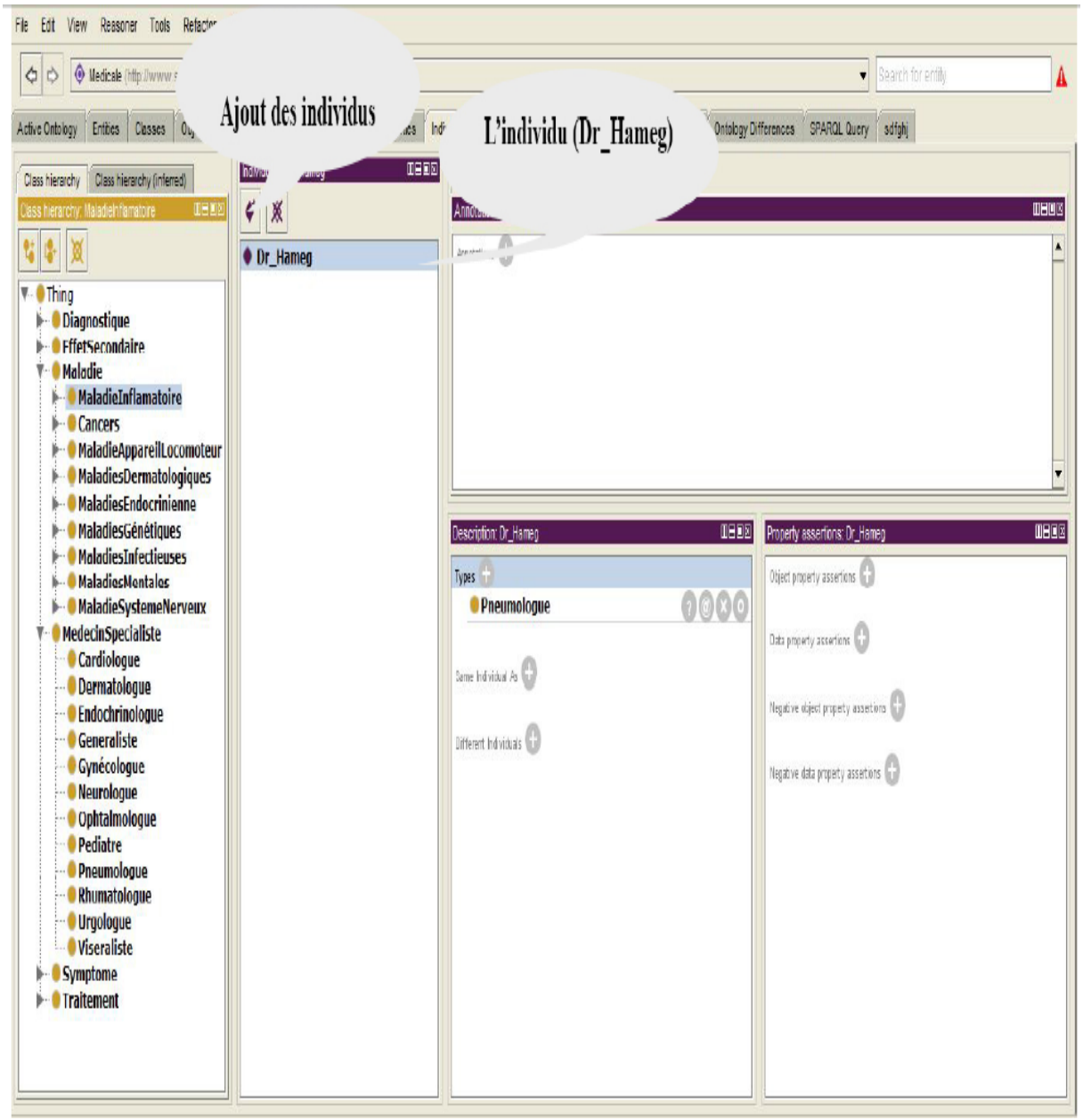


Figure 4.6 Création des individus

4.5.7 Création des Axiomes

Les axiomes sont des restrictions sur les classes, les instances ainsi que les propriétés, ils démontrent comment les concepts interagissent les uns par rapport aux autres et permettent d'établir des inférences. On appelle la classe sur laquelle on restreint la propriété _ filler _.

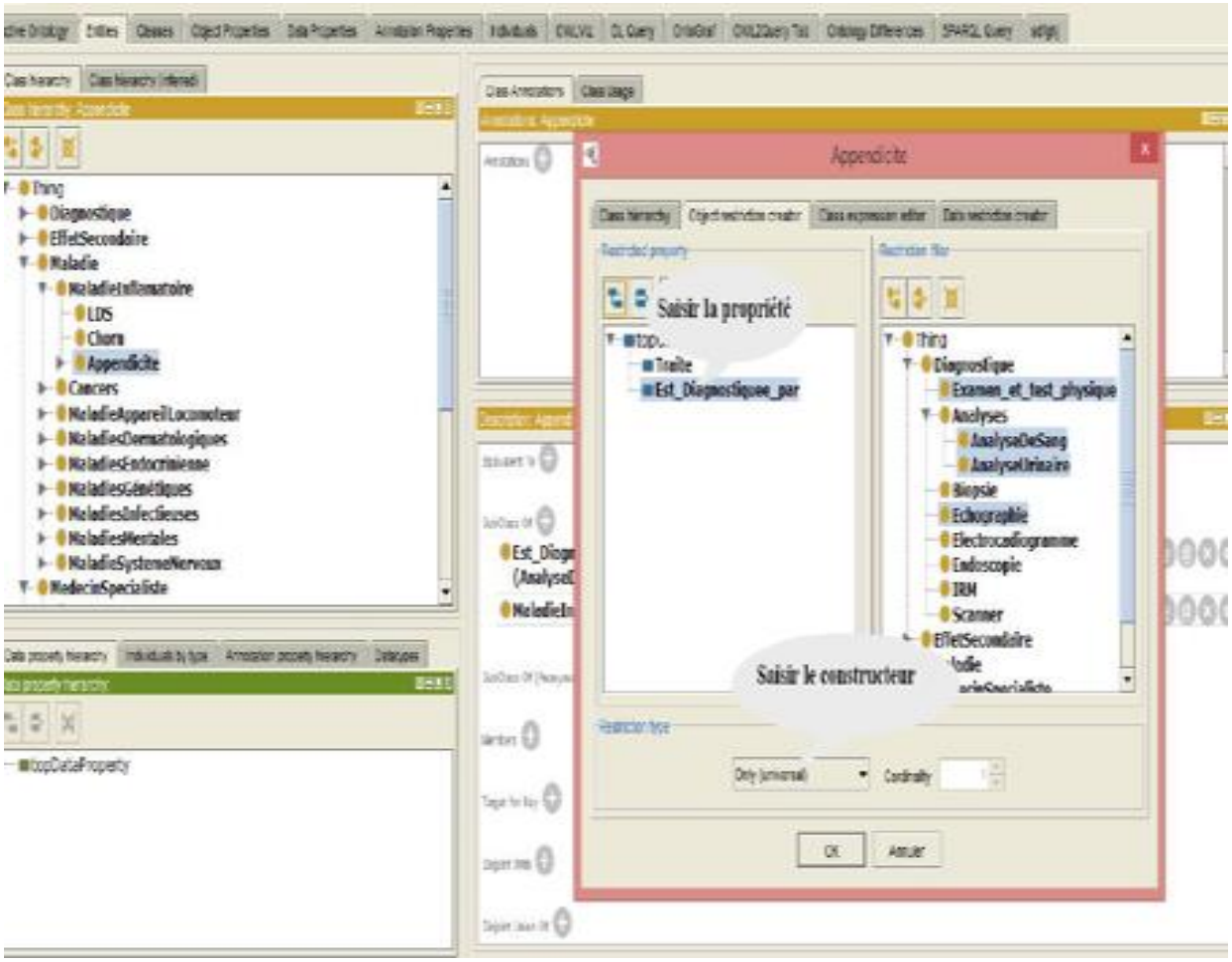


Figure 4.7 Création d'un axiome sur la classe Appendicite

4.5.8 Génération du code RDF/XML

Sous protégé on peut générer plusieurs code dans la barre d'outil on choisit Window puis show Ontology views puis RDF/XML, la figure suivante présente le code généré.

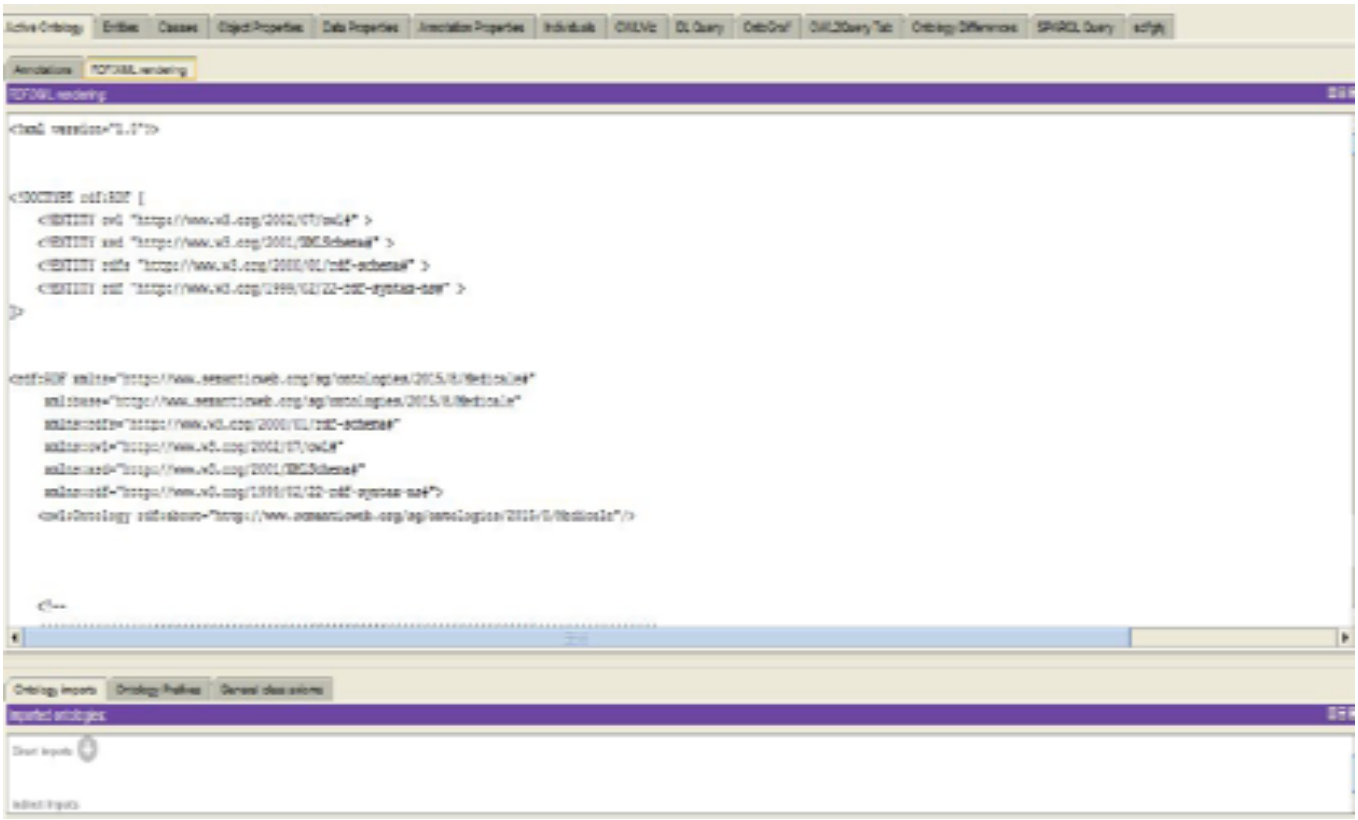
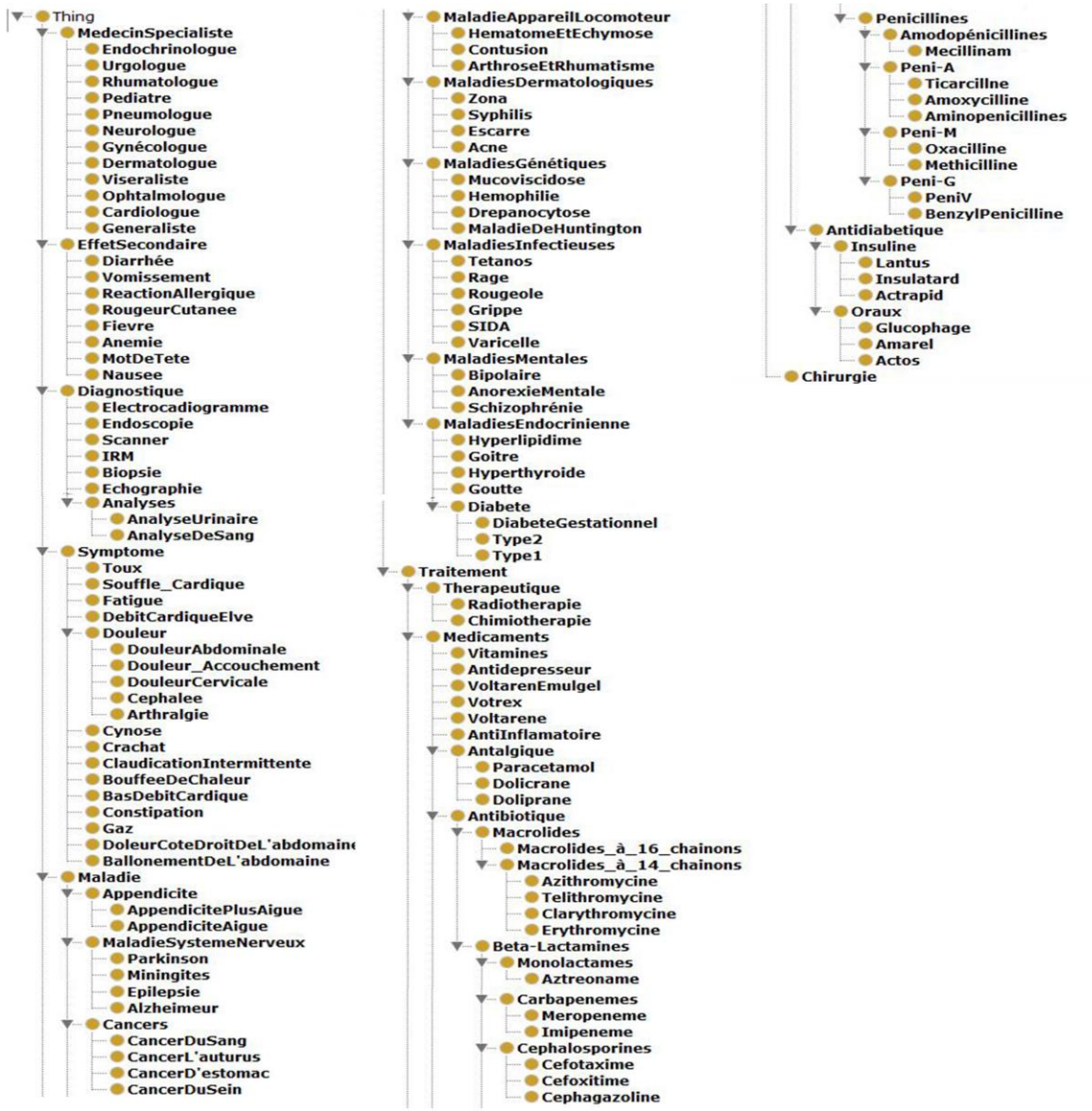


Figure 4.8 extrait de code RDF/XML

4.5.9 Les classes et la hiérarchie des classes de notre ontologie



Les classes et la hiérarchie des classes de notre ontologie

4.6 Etapes de construction de site

L’utilisateur accèdes directement au moteur de recherche, voir l’interface ci-dessous :

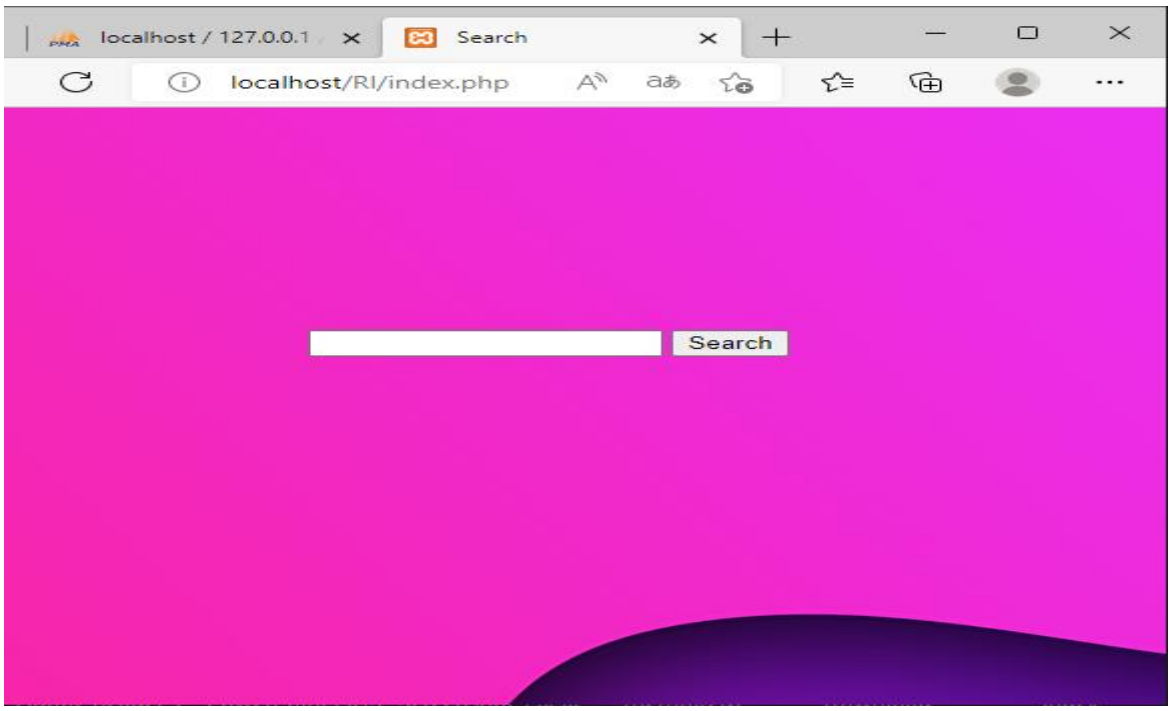


Figure 4.9 interface moteur de recherche

- La figure suivante qui montre la recherche :

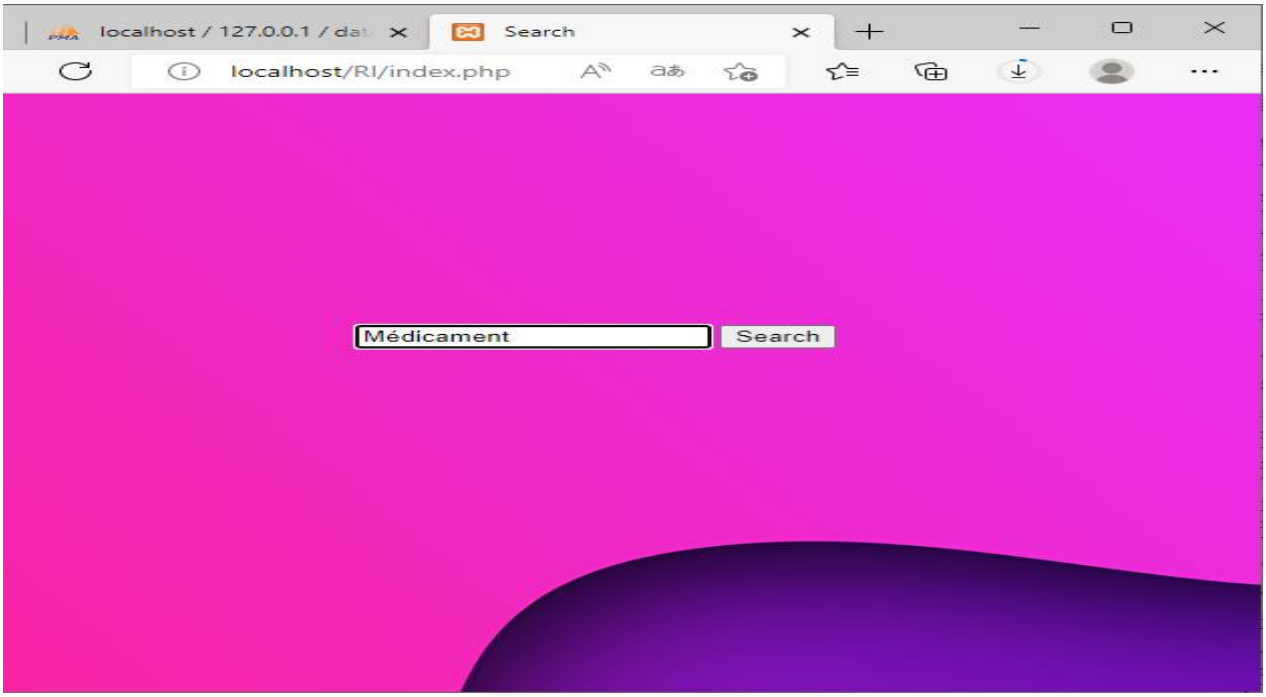


Figure 4.10: Faire la recherche

- La figure suivante qui montre le résultat de la recherche :

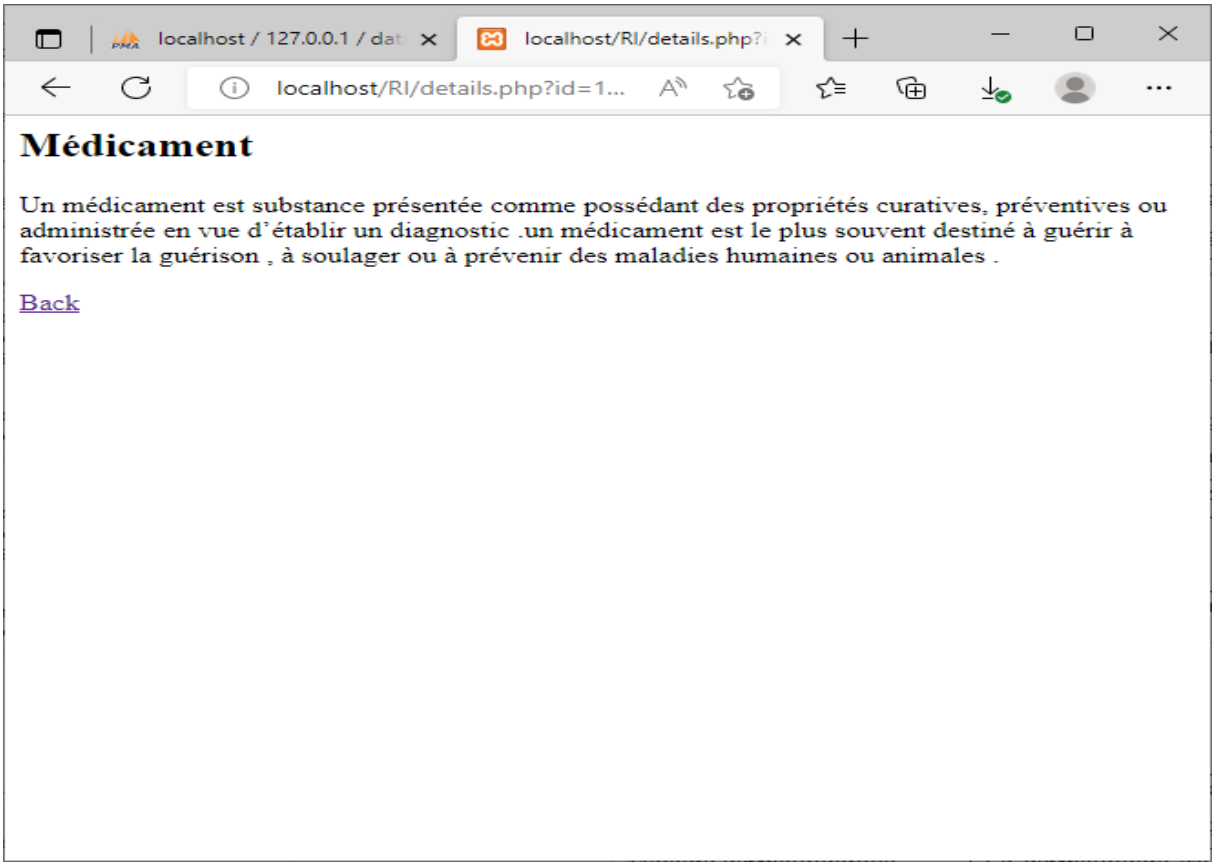


Figure 4.11 le résultat de la recherche

- Lorsque des informations saisies n'existent pas dans la base de données, elles apparaissent « no data found »

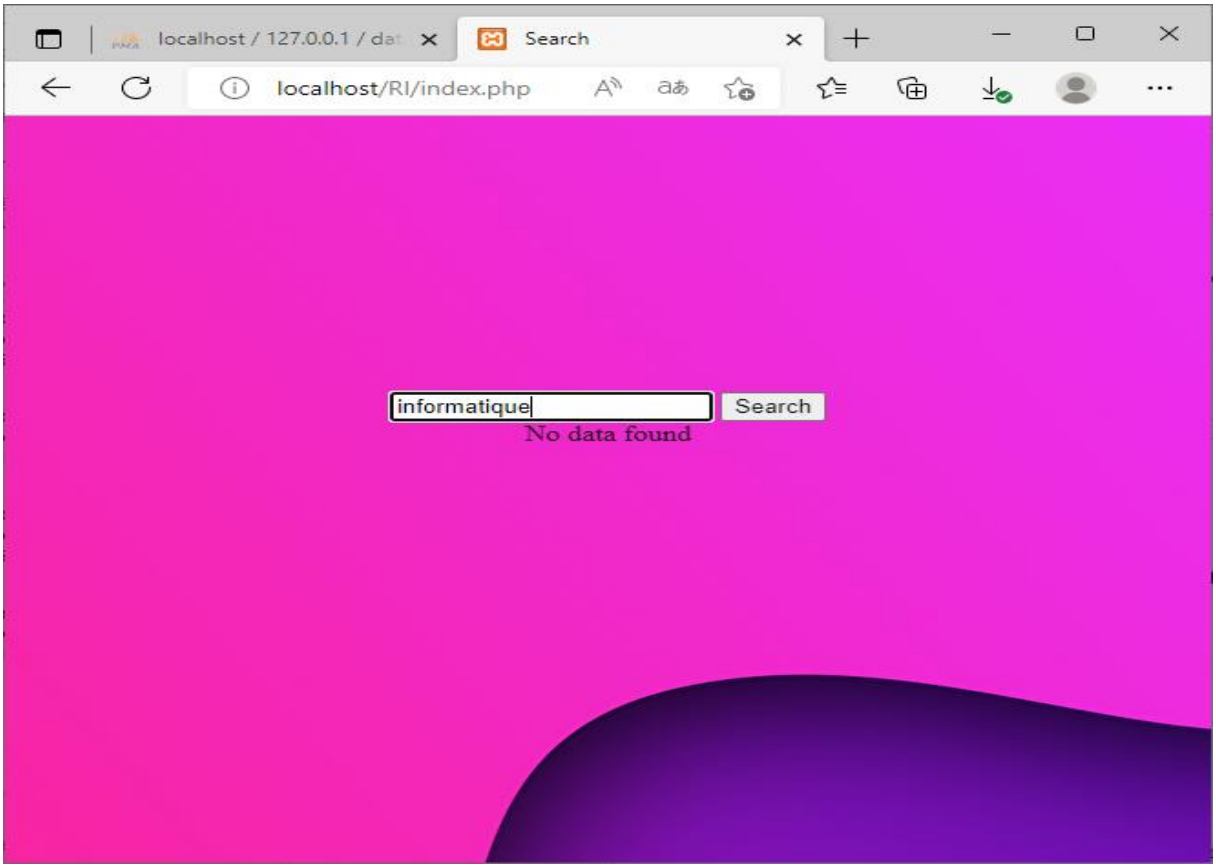


Figure 4.12 afficher message « no data found »

Conclusion

Nous venons de réaliser un système de recherche d'information basé sur d'ontologie.

Ontologie Owl pour l'indexation et la recherche des informations.

Conclusion générale

L'objectif principal de notre système est la mise en place d'un système de recherche des informations en utilisant le Language Owl.

L'extraction manuelle de la connaissance à partir de corpus est un processus très couteux en temps. Il n'a pas toujours été facile de déterminer si un terme devait être représenté en tant que classe ou attribut et certaines connaissances du domaine. Nous nous sommes basés sur les critères de T. Gruber pour évaluer notre ontologie. Les recherches dans le domaine médical ne cessent d'évoluer surtout en matière de nouvelles techniques d'imagerie et de nouveaux médicaments, c'est pour cela qu'il est nécessaire de prévoir une possibilité d'évolution pour l'ontologie.

Notre ontologie évolue par l'ajout manuel de nouveaux concepts il suffit de déterminer à quel axe primitif appartiennent ces nouveaux concepts.

Chaque chapitre de notre mémoire occupe une place importante :

- **Chapitre 1 :** est consacré à la recherche d'information sur le web.
- **Chapitre 2 :** En ingénierie ontologique (qu'est-ce qu'une ontologie, méthodes de construction d'ontologies, processus de construction d'ontologies ...)
- **Chapitre 3 :** parle de la conception.
  Utiliser l'ontologie pour faire de la recherche d'informations en médecine. Et étendre cette ontologie par l'ajout de nouveaux concepts, pour en faire une
- **Chapitre4 :** expliquer les différentes étapes de la mise en œuvre de notre site.

Bibliographies

[1] Cutts, M. Spotlight keynote. In Proceedings of Search Engine Strategies, 2012

[2] Lechani, L., Boughanem, M. Accès personnalisé à l'information : Approches et techniques. Rapport interne, IRIT, 2005.

[3] Adar, E., Teevan, J., Dumais, S.T. & Elsas, J.L. The Web changes everything: Understanding the dynamics of web content. Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, pp. 282–291, 2009.

[4] Cutler, M. Shih, Y. Meng, W. Using the Structure of HTML Documents to Improve Retrieval. Proceedings of the USENIX Symposium on Internet Technologies and Systems Monterey, pp. 22-22 1997.

[5] Collins-Thompson, K. Ogilvie, P. Zhang, Y. Callan, J. Information Filtering, Novelty Detection, and Named Page Finding. TREC-11 Notebook Proceedings, 2002.

[6] Ogilvie, P., Callan, J. Combining document representations for known-item search. Proceedings of *ACM SIGIR conference on Research and development in information retrieval*, pp. 143–150, 2003.

[7] Ogilvie, P., Callan, J. Combining structural information and the use of priors in mixed named-page and homepage finding. *TREC-12 Notebook Proceedings* (Gaithersburg, MD, USA, November 2003), NIST.

[8] Agosti, M., Crivellari, F., Melucci, M. The effectiveness of meta-data and other content descriptive data in
web information retrieval. *Proceedings of the Third IEEE Meta-Data Conference*, Bethesda MD, pp. 139-
149, 1999.-

[9] Zhang, J., Dimitroff, A. The impact of metadata implementation on webpage visibility in search engine results (Part II). *Information Processing and Management*, 41(3):pp. 691–715, 2005.

[10] .Kraaij, W., Westerveld, D., Hiemstra, D. The Importance of Prior Probabilities for Entry Page Search. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27–
34, 2002.

[11] Kamps, J., Mishne, G., de Rijke, M. Language Models for Searching in Web Corpora, 2005.

[12] Ho, J., Garcia-Molina, H., Page, L. The anatomy of efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7), pp. 161–172, 1998.

[13] Chakrabarti, S., Dom, B. Indyk, P. Enhanced hypertext categorization using hyperlinks. In L. M. Haas and A. Tiwary, editors. *Proceedings ACM SIGMOD International Conference on Management of Data*, pp. 307– 318, 1998.

[14]. Bharat, K., Broder, A., Dean, J., Henzinger, M.R. A comparison of techniques to find mirrored hosts on the www. *IEEE Data Engineering Bulletin*, 23(4), pp. 21–26, 2000.

[15]. Dean, J., Henzinger, M. R. Finding related pages in the World Wide Web. *Computer Networks*, 31(11-16): pp. 1467–1479, 1999.

[16] Craswell, N., Hawking, D., Robertson, S. Effective Site Finding using Link Anchor Information, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 250-257, 2001..

[17] Zhu, X. L., Gauch, S. Incorporating quality metrics in centralized / distributed information retrieval on the World Wide Web. *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, pp. 288–295. 2000.

[18]. Lempel, R., Moran, S. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. *Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking*, pp. 387-401, 2000.

[19]. Crestani, F., Lee, P.L. Searching the web by constrained spreading activation. *Information Processing and*
Management, vol. 36, pp.585–605, 2000.

[20] Savoy, J., Picard, J. Retrieval effectiveness on the web. *Information Processing and Management*, vol. 37, pp. 543–569, 2001.

[21 Oliver, A., McBryan. Genvl and WWWW: Tools for taming the Web. *Proceedings of the First International Conference on the World Wide Web*, Geneva, Switzerland, 1994.

[22] Eiron, N., McCurley, K.S. Analysis of Anchor Text for Web Search. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 459-460, 2003.

[23] Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J. Automatic resource list compilation by analyzing hyperlink structure and associated text. *Proceedings of the 7th International World*
Wide Web Conference, pp. 65-74, 1998.

[24] Glover, E.J., Tsiouliklis, K., Lawrence, S., Pennock, D.M., Flake, G.W. Using Web Structure for Classifying and Describing Web Pages. *Proceedings of the 11th international conference on World Wide Web*, pp. 562- 569, 2002.

[25] Diaz, F., Jones, R. Using temporal profiles of queries for precision prediction. *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 18–24, 2004.

[26] Li, X., Croft, W.B. Time-based language models. *Proceedings of the twelfth international conference on*
Information and knowledge management, pp. 469–475, 2003.

[27] Croft, W.B. Combining approaches to information retrieval. In Croft, W. B. (Ed.), Advances in Information Retrieval: Recent Research from the Centre for Intelligent Information Retrieval, *Kluwer Academic*
Publishers, pp.1-36, 2002g regression. Proceedings of the 19th Conference on Learning Theory, pp. 605-619, 2006.

[28] Aslam, J. A., Montague, M. Models for metasearch. *Proceedings of the 24th annual int. ACM SIGIR conf. On Research and development in information retrieval*, pp. 276–284, 2001.

[29] Lee, J.H. Analyse of multiple evidence combination. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267-176, 1997.

[30] Fox, E.A., Shaw, J.A. Combination of multiple searches. In Harman, D.K. (Ed.), *Proceedings of the 2nd Text Retrieval Conference (TREC-2), NIST Special Publication* 500-215, pp. 243-249, 1994.

[31]. Bartell, B.T., Cottrell, G.W., Belew, R.K. Automatic combination of multiple ranked retrieval systems. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 173 – 181, 1994.

[32] Fox, E.A., Nunn, G., Lee, W. Coefficients for combining concept classes in a collection. *Proceedings of the 11th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–308. 1988.

[33] Tsikrika, T., Lalmas, M. Combining evidence for Web retrieval using the inference network model: an experimental study. *Information Processing and Management*, vol. 40, 2004.

[34] Liu, T.-Y., Xu, J., Qin, T., Xiong, W., Li, H. Letor: Benchmark dataset for research on learning to rank for information retrieval. *LR4IR, in conjunction with SIGIR*, 2007.

[35] .Xu, J., Li, H. Adarank: a boosting algorithm for information retrieval. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA. pp. 391–398, 2007.

[36] Liu, T.Y., Learning to Rank for Information Retrieval, *Springer*, 2011.

[37] Cossock, D., Zhang, T. Subset ranking using regression. *Proceedings of the 19th Conference on Learning Theory*, pp. 605-619, 2006.

[38] Nallapati, R. Discriminative models for information retrieval ». *Proceedings on the 27th annual international SIGIR Conference on Research and Development in Information Retrieval*, pp. 64-71, 2004.

[39] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G. Learning to rank using gradient descent. *Proceedings of the 22nd international conference on Machine learning*, ACM. New
York, NY, USA. pp. 89–96, 2005.

[40] Tsai, M.-F., Liu, T.-Y., Qin, T., Chen, H.-H., Ma, W.-Y. Frank: a ranking method with □delity loss. In 'SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval'. ACM. New York, NY, USA. pp. 383–390, 2007.

[41] Cao Z., Qin T., Liu T.Y., Tsai M.F., Li H., « Learning to rank: From pairwise approach to listwise approach ». *Proceedings of the 24th International Conference on Machine Learning*, pp. 129-136, 2007.

[42] Yue Y., Finley T., Radlinski F., Joachims T. A support vector method for optimizing average precision. *Proceedings of the 30th annual international SIGIR Conference on Research and Development in*
Information Retrieval, pp. 271-278, 2007.

– **[Bachimont 2003]** B. Bachimont, J. Charlet & R. Troncy, Ontologies pour le Web Sémantique. Action spécifique 32 CNRS / STIC Web sémantique Rapport final. 2003.

– **[Borst, 1997]** Borst W. N. Construction of Engineering Ontologies. Center for Telematica and Information Technology, University of Tweenty, Enschede, NL. [M.uschold & M.gruninger, 1996b] Uschold M. ET Gruninger M. ”Ontologies : Principles, Methods and Applications”. (1996b).

– **[Farquhar & Fikes, 1996]** Farquhar (A.), Fikes (R.), Rice (J.),1996 : _The Ontolingua Server: A Tool for Collaborative Ontology Construction_, Proceedings of the 10th Knowledge Acquisition for Knowledge- Based Systems Workshop, Banff, Alberta, Canada, p. 44.1-44.19, 1996.

– **[F. Amourache, 2008]** F.Amourache ”Construction d’une ontologie pour l’annotation des cvs/ offres d’emploi” ; le 01/12/2008 ; Université Mentouri de Constantine ; Algérie.

– **[Guarino, 1997a]** Guarino N. Some organizing principles for a unified top-level ontology. AAAI Spring Symposium on Ontological Engineering, 57-63. (1997a).

– **[Guarino; 1997b]** Guarino N. (1997b). Understanding, building and using ontologies. International J. Human-Computer Studies, 46, 293-310, (1997b).

– **[Gomez-Perez, 1999a]** Gomez-Perez, Ontological Engineering: A state of the art. Expert, 1999.

[Gomez-Perez, 1999b] Gomez-Perez, Tutorial on ontological Engineering,
Paper presented at the proc, 1999.

[G.Falquet ,2001] G. Falquet & C.L. Mottaz-Jiang, ”Navigation hypertexte dans une ontologie multipoints de vue”, N^ımesTIC’. 2001.

– **[Lee 2002]** T.B. Lee et al.,”The semantic Web”. In Scientific American, May 2002.

– **[M. Kifer, G. Lausen and J. Wu, 1995]** M. Kifer, G. Lausen and J. Wu”Logical Foundations of Object- Oriented and Frame-Based Languages”. Journal of the ACM (JACM), 42(4):741? 843.

– **[M.gruninger & M.fox, 1995]** Gruninger, M. & Fox, M.S. Methodology for the Design and Evaluation of Ontologies. Proceedings of the IJCAI-95 Worshop on Basic Ontological Issues in Knowledge Sharing, 1995.

– **[M. Uschold & M. King, 1995]** M. Uschold& M. King,"Towards a me-thontology for building ontologies", in Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCA, 1995.

– **Noy, N. F., Fergerson, R. W., & Musen, M. A. (2000)**. The knowledge model of Protege-2000: Combining interoperability and flexibility. In R. Dieng, & O. Corby (Ed.), 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00) (pp. 17? 32). (Lecture Notes in Artificial Intelligence LNAI 1937) SpringerVerlag.

– **[Psyché, Mendes & Bourdeau, 2004]** Valéry PSYCH´E, Olavo MENDES,
Jacqueline BOURDEAU, Apport de l ?ingénierie ontologique aux environnements de formation à distance, 2004.

– **[Tom Gruber, 2009]** Tom Gruber dans l'Encyclopédie des systèmes de base de données, Ling Liu et M. Tamer Ozsu (Eds.), Springer-Verlag, 2009.

– **[T. Berners-Lee, Hendler & O. Lassila, 2001]** T. Berners-Lee, J. Hendler and O. Lassila, (2001). "The SemanticWeb". Scientific American, 284(5):34? 43.

– **[Oberle 2004]** D. Oberle, R. Volz, B. Motik ET S. Staab,"An extensible ontology software environment". In S.Staab ET R. Studer (Eds.), Handbook on Ontologies (pp. 299-320): Springer Verlag. 2004.

– **[Uschold 2002]** M. Uschold and M.Gruninger, «Creating semantically integrated communities on the World Wide Web". Honolulu: Semantic Web
Workshop, 2002.

Références Webiographique

– **[43] :** http : //www.cartographie-semantique.fr/propositions/sndf-notre-formalismede-description-pour-la-cartographie-semantique

– **[44] :** http : //www.lehtml.com/xml/syntaxe.html

– **[45]:** http: //www.w3schools.com/webservices/ws RDF intro

– **[46]:** http: //www.w3schools.com/webservices/ws RDF schema

– **[47]:** http: //www.w3.org/2004/OWL/

– **[48]:** http: //www.w3.org/2004/Daml-Oil/

– **[49]:** http: //www.w3.org/2001/sw/WebOnt/

/